TECHNICAL REPORT

Additional Analysis of the ESTCP Discrimination Study Data at Camp Sibert, Alabama

ESTCP Project MM-0504

OCTOBER 2008

Dr. Stephen Billings
Dr. Leonard Pasion
Sky Research, Inc)
Laurens Beran
UBC-GIF

Environmental Security Technology Certification Program

Sky Research Inc.

Additional Analysis of the ESTCP Discrimination Study data at Camp Sibert, AL

Project 200504: Practical Discrimination Strategies for Application to Live Sites

Submittal Date: October 20, 2008
Version 2.0

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	10/20/2008	January 2008-October 2008

4. TITLE AND SUBTITLE
Addendum to Camp Sibert Data Processing Demonstration Report
Project 200504: Practical Discrimination Strategies for Application to Live Sites

5. FUNDING NUMBERS
Contract # W912HQ-05-C-0018

6. AUTHOR(S)
Dr. Stephen Billings, Dr. Leonard Pasion (Sky Research, Inc.)
Laurens Beran (UBC-GIF)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Sky Research, Inc.
445 Dead Indian Memorial Rd.
Ashland, OR 97520

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ESTCP Program Office
901 North Stuart Street, Ste 303
Arlington, VA 22203-1821

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified/Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Geonics EM61 cart, MTADS EM61 and EM-63 cart data from Camp Sibert were investigated to determine if the number of "can't analyze" anomalies could be reduced and more objective stop-digging criteria selected. Many of the "can't analyze" anomalies in the MTADS EM61 were caused by cart-bounce along North-South transects and could be avoided by using only East-West transect data. However, poor data coverage increased the chances of generating false negative declarations. If North-South transects are retained, sensor motion relative to the ground can cause difficulties in obtaining good model fits to the data. A modeling framework developed elsewhere was used to determine if an anomaly was caused by sensor motion or a compact metallic target. The method has promise but had limited applicability due to the lack of accurate ground-clearance and topographic data. To determine an objective operating point, the training data must be representative of the test-data. In particular, outliers need to be avoided and here this was achieved by using multiple feature vectors, obtained by analysis of a depth versus misfit curve, in the classification. Using this technique a false-negative in the EM-63 data could be avoided.

14. SUBJECT TERMS

15. NUMBER OF PAGES
42

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Executive Summary

For the Camp Sibert discrimination study, the team of Sky Research and the University of Brish Columbia-Geophysical Inversion Facility (UBC-GIF) created 8 different dig-sheets from 6 different sensor combinations: (i) MTADS magnetics; (ii) EM61 cart (classification and size based); (iii) MTADS EM61 (classification and size based); (iv) MTADS EM61 and magnetics; (v) EM63; and (vi) EM63 and magnetics. Effective discrimination was demonstrated for all sensor combinations with just one false-negative for the EM63 when inverted without magnetometer location constraints. The EM61 cart, MTADS EM61 and MTADS EM61 interpretations suffered from large numbers of "can't analyze" anomalies that had to be excavated because an accurate model fit could not be obtained. A large number of these can't analyze anomalies were from geologic sources. For the MTADS EM61 these "geological anomalies" were caused by spurious signal generated from cart bounce on North-South transects. A further problem with the MTADS EM61 and EM61 cart datasets was the conservative stop-dig threshold which resulted in the excavation of many non-hazardous items.

In this report we further analyze the EM-61 cart, MTADS EM-61 and EM-63 datasets collected at Camp Sibert. In particular we investigated methods to reduce the number of "can't analyze" anomalies in the EM-61 cart and MTADS EM-61 datasets, and methods for objectively setting the stop-digging threshold in all datasets.

We found that the number of "can't analyze" anomalies in the MTADS EM-61 data could have been significantly reduced by monitoring the amplitudes or signal-to-noise ratio of data collected along North-South and/or East-West transects. Simply by rejecting anomalies that did not exceed the picking threshold of 25 mV on an E-W transect would have resulted in a reduction of 134 can't analyze anomalies in the MTADS EM-61 dataset (from a total of 285) and 88 can't analyze anomalies when the MTADS EM-61 data were cooperatively inverted (from 205). Similar reductions could be obtained by applying a metal-geology pre-screener which used the relative SNR in N-S and E-W to determine the likelihood of metal.

We investigated the cause of the can't analyze anomalies of geological origin whose amplitude exceeded 25 mV along the E-W transects. We hypothesized that many of these anomalies were due to sensor movement relative to ground. The background response was modeled by estimating a ground clearance from the elevation data and assuming that the background magnetic susceptibility was uniform in each cell. The accuracy of our modeling was limited due to filtering artifacts in the observed data, the accuracy of the ground clearance estimate, and small scale topography (i.e. depressions and bumps on the surface that would affect the measured data).

In many cases we found that small scale anomalies could be predicted using our modeling techniques and that sensor movement was indeed the likely cause. We also found that there were a number of metallic anomalies whose response could not be properly modeled due to variations in the signal from the background due to sensor movement. There were also several anomalies that were caused by longer wavelength spatial variations in magnetic soil properties that were not suppressed by the detrend filters that were used to pre-process the sensor data. We conclude that sensor movement relative to the ground is an important contributor to false-alarms and that we, in principal, have techniques to prevent such false declarations. However, the lack of good ground-clearance and micro-topographic information prevents the effective use of these techniques.

We also investigated discrimination performance when polarization tensor models were obtained from East-West transects only. When depths are constrained by cooperative inversion, there was very little difference in discrimination performance when using all lines or only East-West lines. With all data, only 2 false-positives were required before all 118 UXO were excavated compared to 4 false-positives for the East-West only data. When the cooperative constraints were not used there was one UXO that was recovered quite late in the dig-list, just 4 excavations before the operating point. Poor spatial coverage was deemed to be a possible cause for the poor fit obtained for that particular anomaly. We conclude, that, at least where data coverage was acceptable, there was very little benefit gained by collecting the MTADS data along perpendicular traverses.

At Camp Sibert, the stop-digging points were selected intuitively based on the characteristics of the training data. The thresholds were set very conservatively due to the potential for “outliers” (UXO with feature vectors that differ significantly from the training data). We addressed the outlier issue by using multiple feature vectors for each anomaly. Performance was improved for the EM-63 (the false-negative was prevented) but not for the EM-61 cart data. The performance of the EM-61 cart was degraded because many of the clutter items had relatively poor SNR and had larger, deeper, UXO-like models that fit the data relatively well. This caused the false-alarm rate to increase. The use of multiple feature vectors did prevent the occurrence of outliers and allowed us to objectively set the operating point based on a boot-strap analysis of the training data. For the EM-63 the boot-strap analysis caused very little change in the operating point, while for the EM-61 cart, the multi-feature vector operating point could be set more aggressively. With this more aggressive cut-off the EM-61 cart performance was only slightly worse than was reported in the original demonstration report, with the added advantage that the operating point was based on objective criteria. Any bootstrap analysis of the test-dataset will only be relevant to the training data if the test dataset is representative of the training dataset. Consequently, we developed a technique to determine the statistical similarity of the test and training datasets and used it show that both the EM61 and EM63 training datasets were representative of the test datasets.

1 Introduction

For the Camp Sibert discrimination study, the team of Sky Research and UBC-GIF created 8 different dig-sheets from 6 different sensor combinations: (i) MTADS magnetics; (ii) EM61 cart (classification and size based); (iii) MTADS EM61 (classification and size based); (iv) MTADS EM61 and magnetics; (v) EM63; and (vi) EM63 and magnetics. Effective discrimination was demonstrated for all sensor combinations (Figure 1), with just one false-negative for the EM63 when inverted without magnetometer location constraints. The cued-interrogation EM63 data when cooperatively inverted with the magnetics was the most effective discriminator. The magnetometer had the "worst" inherent discrimination ability as indicated by the high percentage of base-plates and partial rounds that needed to be excavated by the time all the UXO items were recovered (Figure 2). In contrast, the EM61 cart, MTADS EM61 and MTADS EM61 cooperatively inverted datasets required many fewer false-positive excavations to recover all UXO (Figures 1 & 2). However, each of these sensor combinations suffered from a significantly higher number of "can't analyze" anomalies that had to be excavated because they could not be classified. A large number of these can't analyze anomalies were from geologic sources (Figure 3). For the MTADS EM61 these "geological anomalies" were caused by spurious signal generated from cart bounce on North-South transects (Figure 4). A further problem with the MTADS EM61 and EM61 cart datasets was the conservative stop-dig threshold which resulted in the excavation of many non-hazardous items (Figure 2).

To improve discrimination performance the following two issues need to be addressed:

- Reduction of the number of "can't analyze" anomalies, particularly for the MTADS EM61;
- More intelligent selection of the stop-digging point.

In an effort to address these issues we conducted the following additional analysis on the data:

1) Re-analyses of MTADS EM61 data using E-W transects only. By eliminating the N-S transects we can significantly reduce the number of anomalies with a geological origin that exceed the picking threshold of 25 mV. However, with just one set of transects, there is very little excitation of the polarization components in the direction perpendicular to the transect paths. We reinverted anomalies with the N-S transects removed to determine if the inversion fits were degraded;
2) Exploration of principled methods to reduce the number of anomalies that were classified as can't decide. There were three main types of anomalies classified as can't analyze: geological anomalies that provide poor fits to the dipole model; small or deep anomalies where the signal-to-noise ratio is insufficient to constrain the dipole model; and anomalies with data quality issues (poor coverage, bad data etc). We concentrate on analyzing the geological anomalies here. Our concept of a Figure of Merit provided one potential method to remove the can't analyze category and Receiver Operating Characteristic (ROC) curves incorporating this concept were provided in the main demonstration report;
3) Examination of the impact of digging (i.e. incorporating the increase in truth data) on where one can stop digging and estimates of probability that no remaining UXO are present. At Sibert, the stop-digging points were selected intuitively based on the characteristics of the training data. The thresholds were set very conservatively due to the potential for "outliers" (UXO with feature vectors that differ significantly from the training data). Here we attack the outlier issue head-on by using multiple feature vectors for each anomaly. We also attempt to update the underlying distributions and corresponding stop-dig thresholds iteratively as additional labels are revealed during excavations; and

4) In addition to these three main activities, we also reinverted all EM63 data after discovering a problem with the detrending algorithm that was applied to the data.

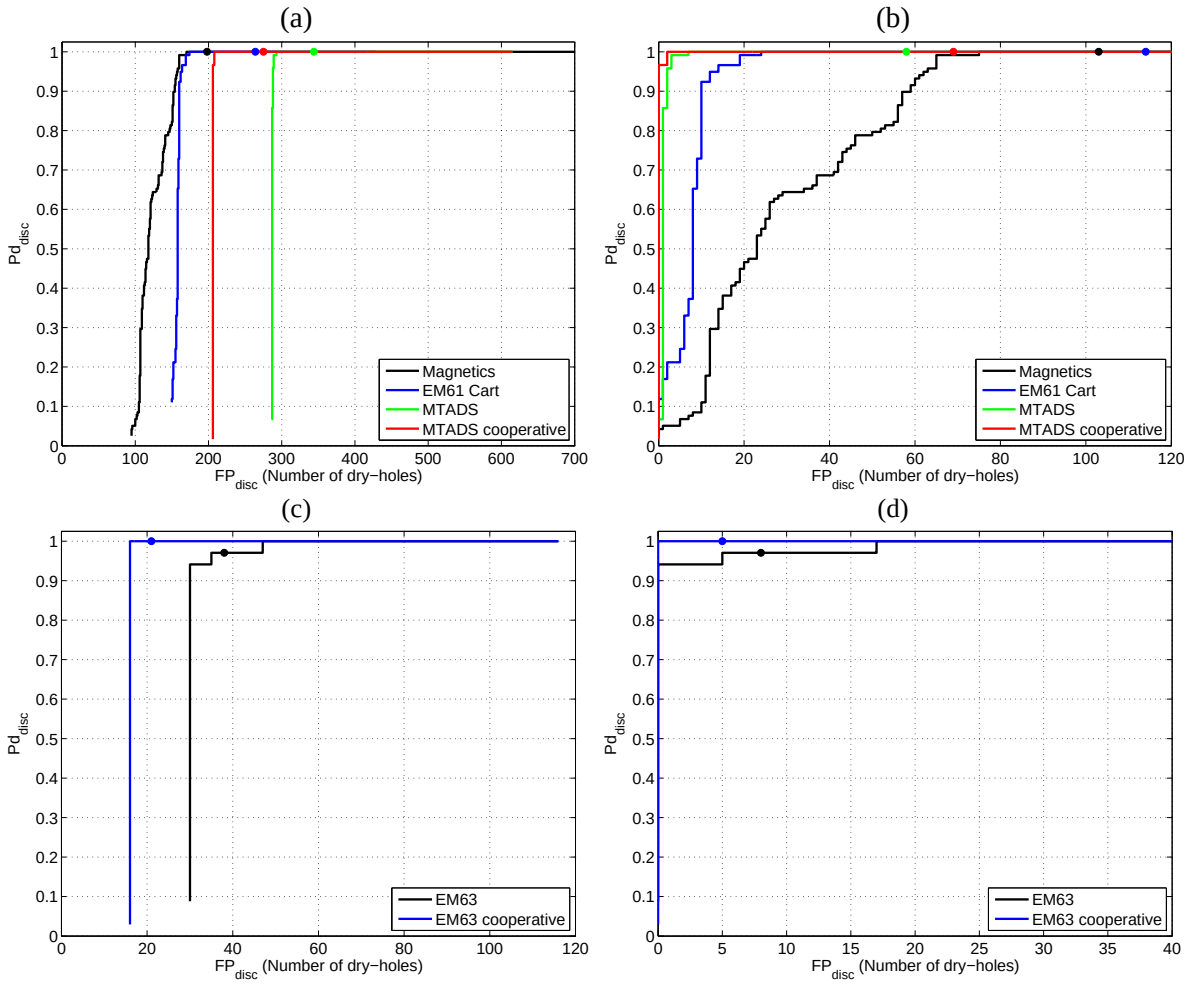


Figure 1: Receiver operating curves for the magnetics, EM61 cart, MTADS and MTADS cooperative inverted datasets (a and b) and the EM63 and EM63 cooperatively inverted (c and d). The graphs in (a) and (c) include the "can't analyze" category", while the graphs in (b) and (d) exclude them. The colored dots represent the stop-digging point for each dataset.

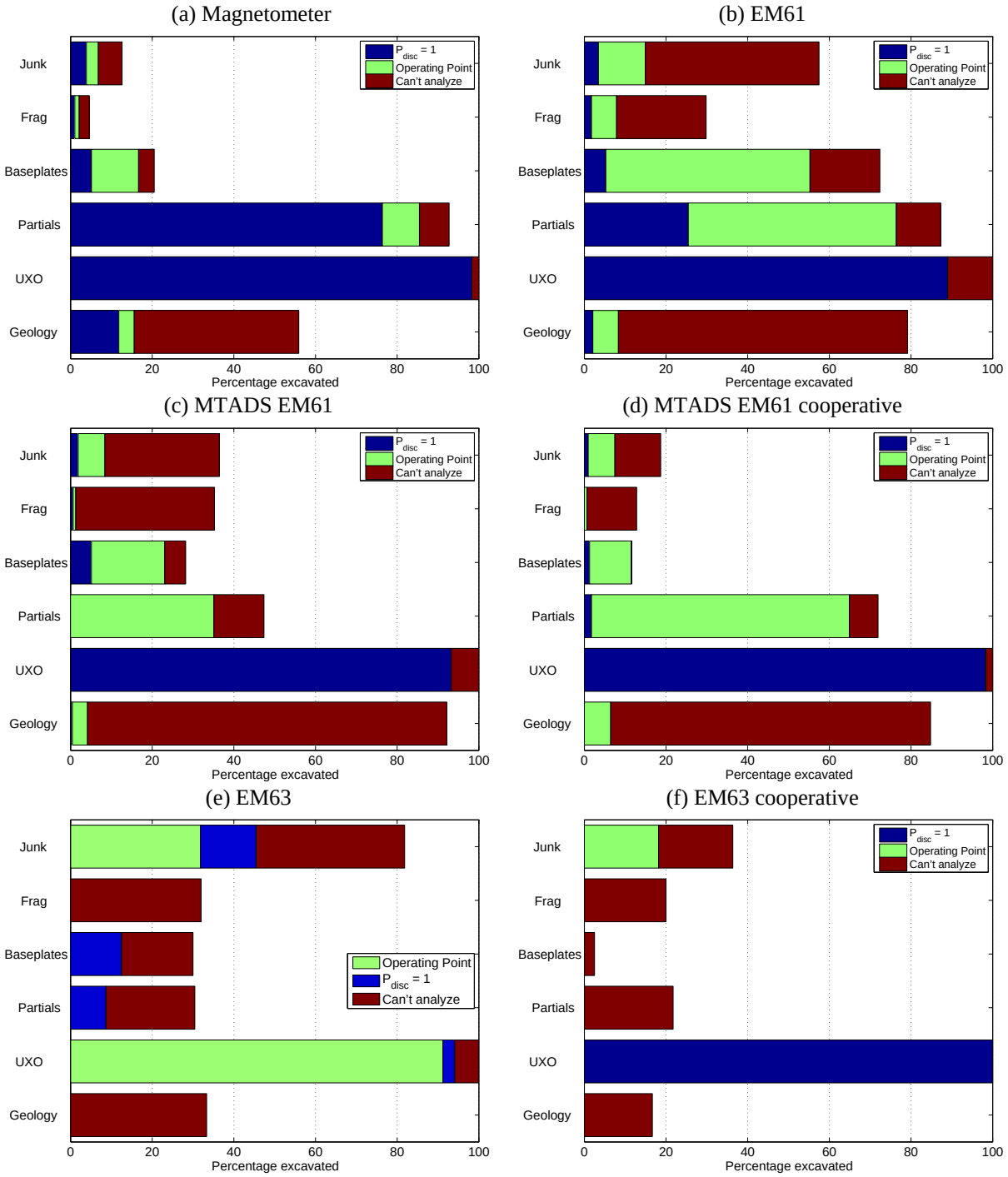


Figure 2: Comparison of the percentage of items of each class that were excavated at the classifier operating point, at the point where Pdisc = 1 and as "can't analyze". Note that the EM63 has the order of the "Operating Point" and "P_{disc}=1" categories reversed as not all UXO were recovered at the operating point.

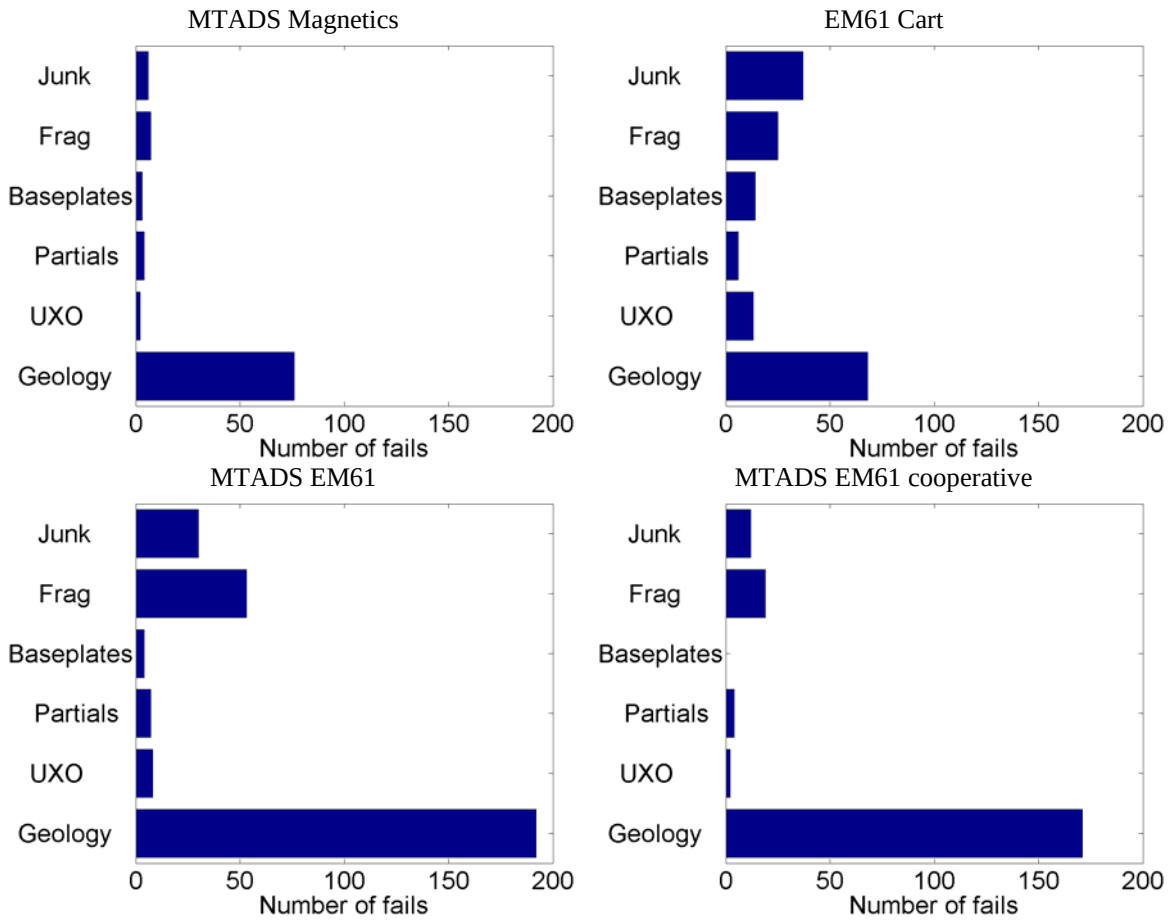


Figure 3: Number of "can't analyze" anomalies in the magnetometer, EM61 cart, MTADS and MTADS cooperatively inverted datasets.

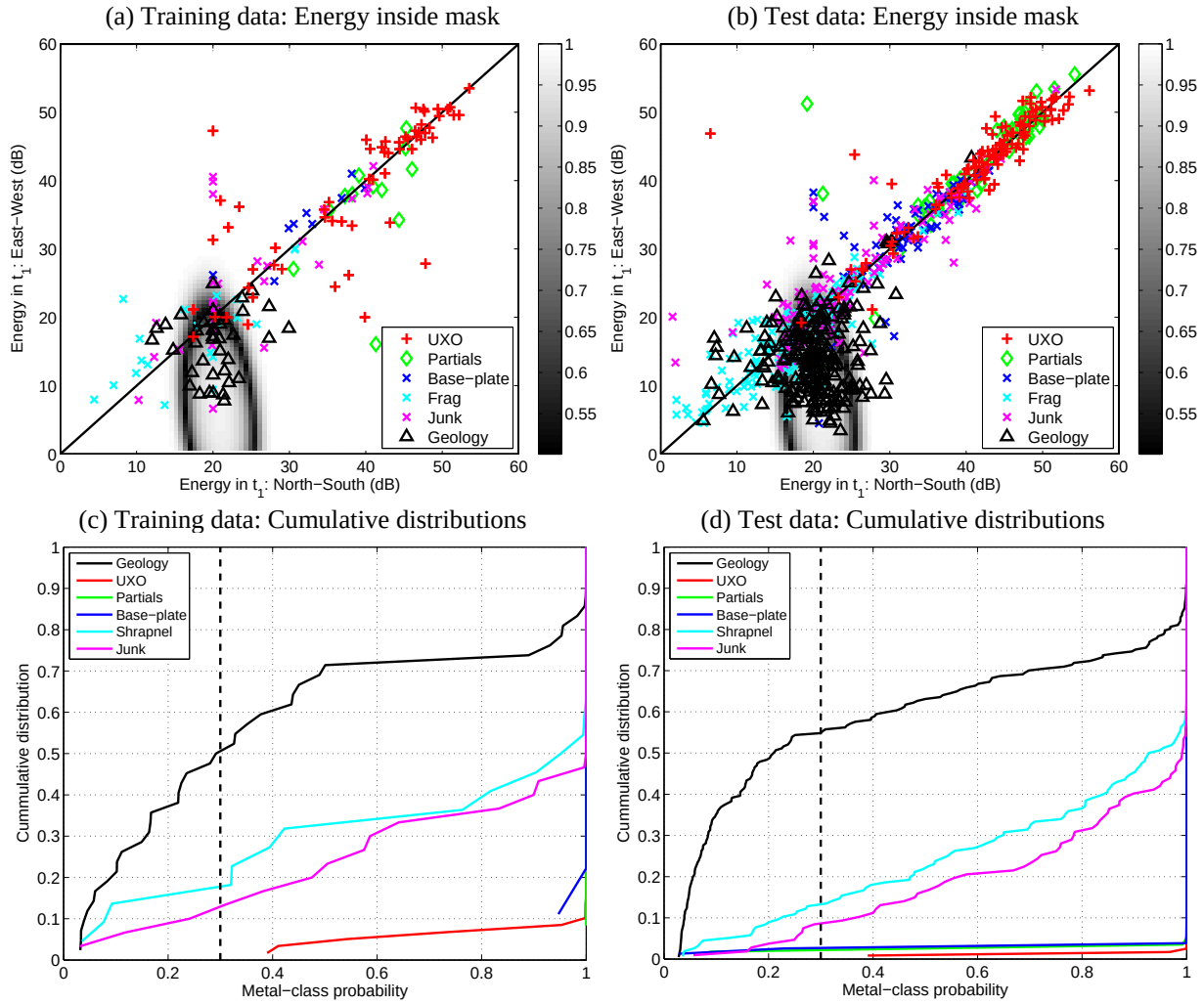


Figure 4: Reduction in geological false-alarms in the MTADS data using the difference in the energy in the North-South versus East-West lines: Energy feature vectors from the (a) training and (b) test-data overlying the classifier decision surface; and cumulative distributions of classes on the (c) training and (d) test-data when ranked by "metal" probability. The vertical dashed lines in (c) and (d) represent the suggested operating point of 0.3 metal-class probability. The UXO, partial and base-plate classes have a significant portion of anomalies with metal-class probability approximately equal to 1. Hence their cumulative distributions increase almost vertically at an abscissa of 1.

2 Reanalysis of MTADS EM61 data using only the E-W transects

2.1 Methodology

One simple method to reduce the number of "can't analyze" anomalies in the MTADS EM61 data would be to remove the North-South transects (as these were the predominant cause of the geological false alarms: Figure 4). In theory, collecting data on perpendicular transects should result in better excitation and recovery of all three polarization tensor components. As we only used features derived from the primary polarization, this may not impact the performance of the discrimination algorithm applied at Camp Sibert.

For this task we implemented the following procedure:

- After removal of the N-S transects, the amplitudes of each anomaly were recalculated. If the anomaly amplitude did not exceed the detection threshold (25 mV) it was removed from the dig-list;
- The MTADS EM61 and MTADS EM61 cooperative anomalies were inverted using only E-W transects;
- All inverted anomalies were manually reviewed to the determine if they fit within the "can't analyze" category;
- An automated Figure-of-Merit was calculated for each anomaly;
- Anomalies were classified using the same feature vectors and training data as the original MTADS EM61 and MTADS EM61 cooperative datasets. This included the use of different thresholds for the low and high FOM anomalies.

2.2 Results

Including the GPO and all training data there were a total of 908 anomalies in the MTADS EM61 data with maximum amplitude greater than 25 mV. After removing the North-South lines, this number was reduced by 167 to a total of 741 anomalies. 141 of these were training data or GPO, with 600 anomalies in the test-data. Therefore, just by eliminating the North-South lines we can reduce the number of test-data by 134 anomalies.

Table 1 and Figure 5 summarize the results for the MTADS EM61 and MTADS EM61 cooperative for all lines and the East-West lines only. At the operating point, there is a significant reduction in the number of anomalies excavated. The reduction is almost entirely due to "can't analyze" anomalies that do not have to be excavated because they don't exceed the 25 mV detection threshold.

	MTADS EM61				**MTADS EM61 cooperative**			
	# alarms	Can't analyze (CA)	False positives (FP)	FP (excluding CA)	# alarms	Can't analyze (CA)	False positives (FP)	FP (excluding CA)
All lines	734	285	344	59	734	205	275	70
East-West only	600	150	207	57	600	121	187	66
Difference	**134**	**135**	**137**	**2**	**134**	**84**	**88**	**4**

Table 1: Number of anomalies in the MTADS and MTADS cooperative dig-sheets for all lines, and East-West lines only.

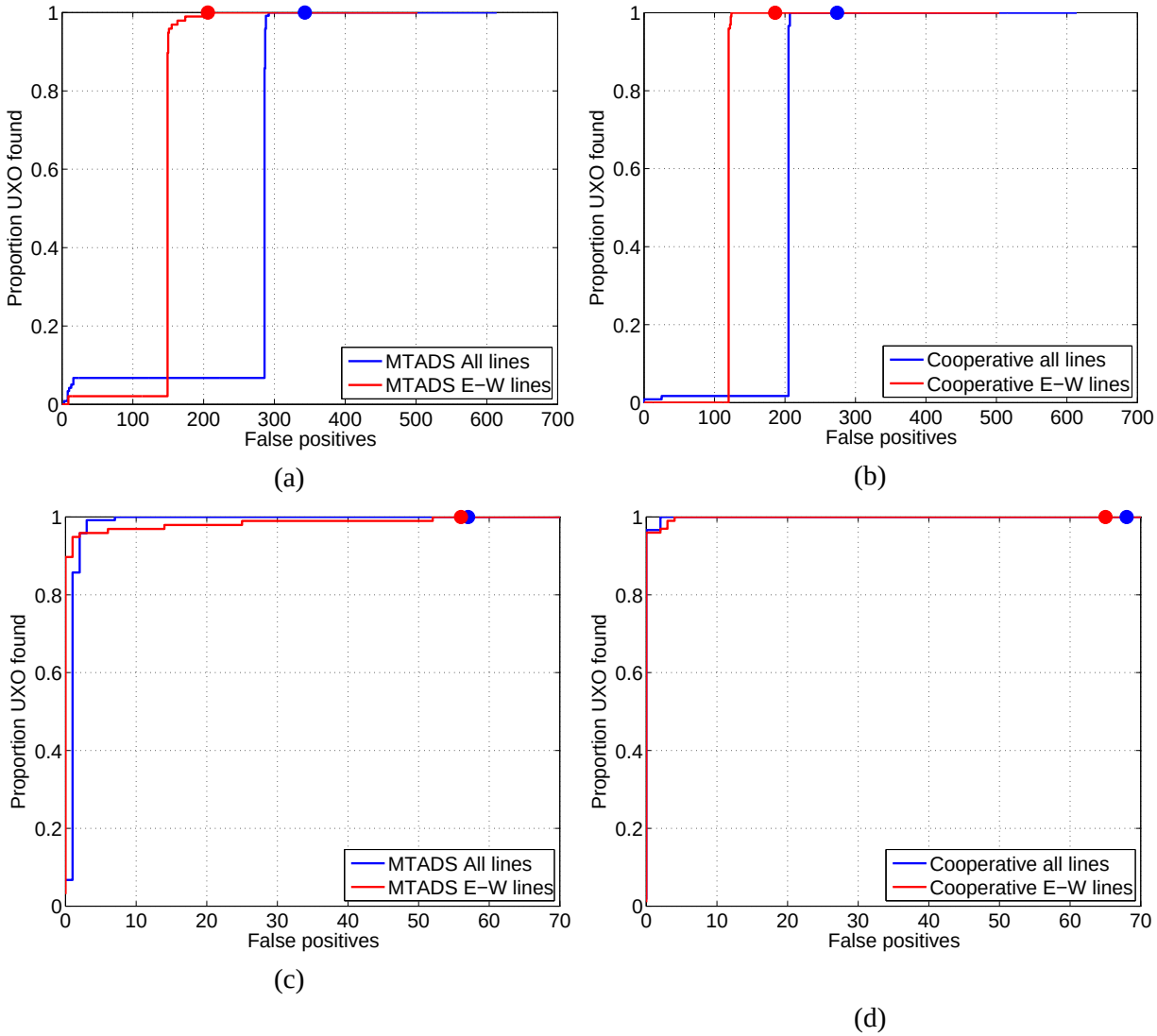


Figure 5: Receiver operating characteristic (ROC) curves for the MTADS EM61 (a and c) and MTADS EM61 cooperative data (b and d). The results are shown including the "can't analyze" category (a and b) and excluding that category (c and d).

When depths are constrained by cooperative inversion, there is very little difference in discrimination performance when using all lines or only East-West lines (Figure 5d). With all data, only 2 false-positives are required before all 118 UXO are excavated compared to 4 false-positives for the East-West only data. When the cooperative constraints are not used (Figure 5c), there is one UXO that is recovered quite late in the dig-list (after 52 false-positives), just 4 excavations before the operating point.

Table 2 lists the key inversion parameters for the last UXO excavated in the East-West only dataset (anomaly number 1302 which was buried at 40 cm depth), while Figures 6 and 7 show the inversion fits for all lines and East-West only lines. The so-called "misfit versus depth" curves in the figures are obtained in the follow way. The optimization routine we use for inversion is a local Newton-type method that minimizes the least squares objective/misfit. We address the problem of local minima and assess the level of ambiguity in resolving the depth of an item by choosing multiple starting models. We start each inversion by scanning the subsurface (x, y, z) up to a 1.2 m depth. At each position we solve for the non-

diagonalized polarization tensor[1] for the first time channel (chosen for its superior signal-to-noise ratio). For each combination of a position and polarization tensor we compute a data misfit (green circles). The depth-misfit curve is defined by the best fit at a given depth (red line). Starting models for the full inversion of multi-channel data are selected along the depth-misfit curve among the models with relative misfit below a given threshold, here 15% (red circles). If the depth-misfit curve contains local minima these are also selected as starting models. These starting models are used to seed 10 full-nonlinear inversions whose final depths and misfits (appropriately scaled) are plotted as black asterisks. The depth versus misfit curve for the inversion with all data has a distinct minimum at 18 cm. Most of the full-non-linear fits cluster at that depth, which is 22 cm shallower than the ground-truth depth. For the East-West only data, the misfit-versus depth curve has two minima, one close to the surface, and the other at 80 cm depth. The full-non-linear solutions cluster near the surface and at 60 cm depth, with the shallower solutions having lower misfits. Because the inversion converges to a model much closer to the surface, the polarization tensor model has a relatively low amplitude which is why it's recovered so late in the dig-list. The poor data coverage of the East-West lines may be one reason why the model depth is unconstrained by the data.

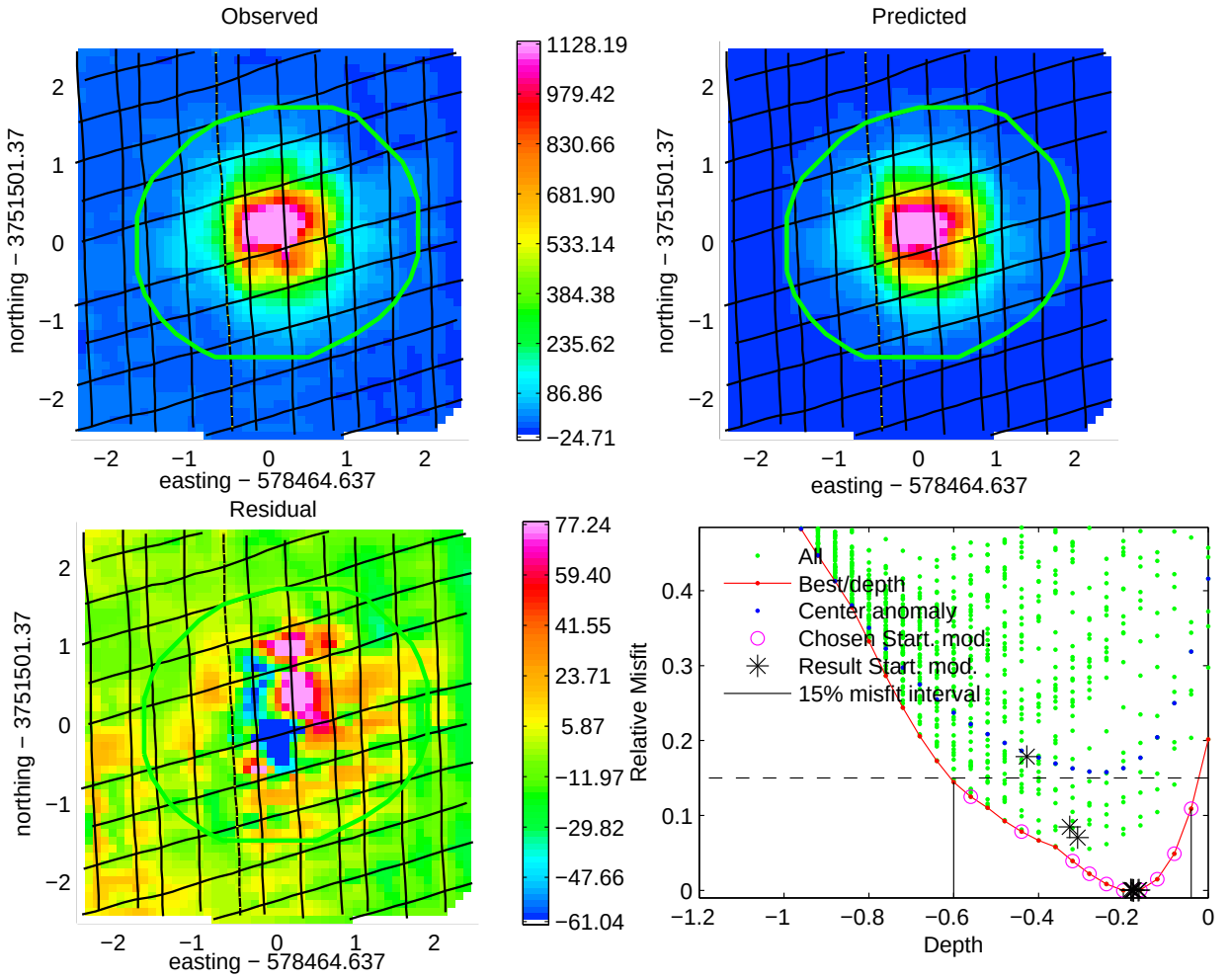


Figure 6: Plan view of time-channel 1 showing MTADS data, predicted data from the recovered polarization tensor model, residual (observed minus predicted) and the misfit versus depth curve.

[1] When the polarization tensor is not explicitly diagonalized, the inverse problem is linear

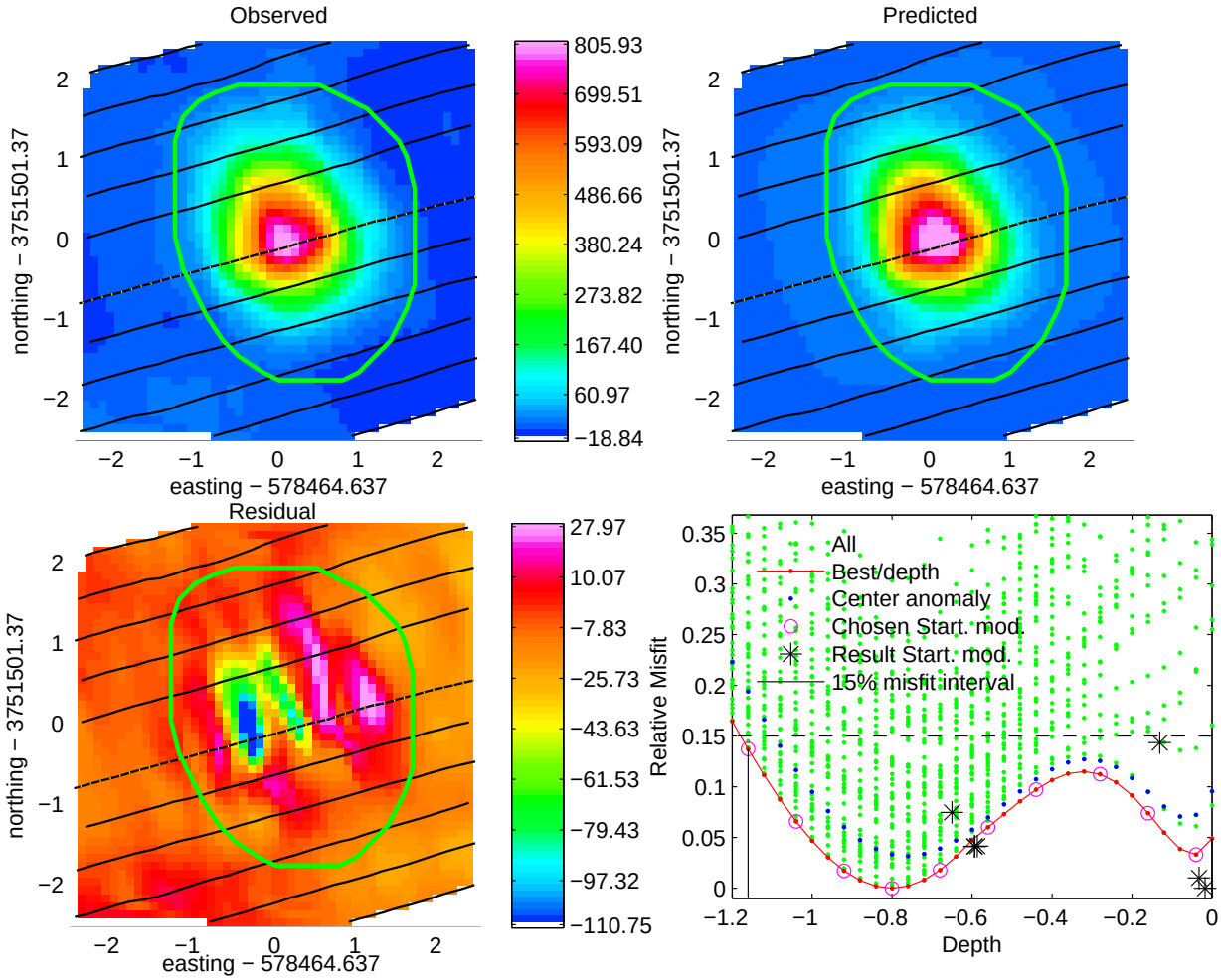


Figure 7: Plan view of time-channel 1 showing MTADS E-W data, predicted data from the recovered polarization tensor model, residual (observed minus predicted) and the misfit versus depth curve.

	Easting error (cm)	**Northing error (cm)**	**Depth error (cm)**	**$L_1(t_1)$**	**$L_1(t_3)$/ $L_1(t_1)$**
All	0.15	0.16	0.22	322	0.294
EW lines only	0.10	0.18	0.38	167	0.204

Table 2. Key inversion parameters for anomaly 1302, the last item excavated with the E-W only data.

3 Exploration of principled methods to handle or reduce the number of anomalies that were classified as can't decide

3.1 Methodology

As outlined in the proceeding sections, the majority of “can’t analyze” anomalies had a geological origin (particularly for the MTADS data). In the demonstration report, we investigated the use of a metal-geology pre-screener that was trained using the anomalies in the ground-truth data (Figure 4). Many of the geological anomalies were clustered in a distinct region of the feature space. In total, the pre-screener could have reduced the number of “can’t analyze” anomalies in the MTADS EM-61 data by 130 (from 285 down to 155). For the MTADS EM-61 cooperative inversion the reduction would have been from 226 down to 116. Similar reductions in “can’t analyze” anomalies would have been achieved by simply rejecting anomalies that don’t meet the threshold criteria on both sets of transect data (Table 1).

For the remaining “can’t analyze” anomalies we have investigated the effectiveness of using a soil model to provide a correction to the MTADS EM-61 data. The background soil response is modeled by (1) representing the transmitting field with a number of dipoles, (2) using the analytic solution of a dipole over a laterally homogeneous earth to calculate the response to each of the dipoles, and (3) summing the response of all the dipoles to approximate the total signal. Specific details of the calculations can be found in the annual report for SERDP 1573 [1]. Although numerous assumptions are used in the calculation, previous analysis of Geonics EM61 MK2 data has shown the technique to be capable of approximating the measured response (Figure 8).

The MTADS array does not have an altimeter to measure the height of the sensor above the ground. Therefore, GPS elevation data are used to estimate a ground clearance. We process the data on a “cell-by-cell” basis, i.e. we analyze a 5 m x 5 m portion of the total data set centered on each target anomaly. We assume that within each cell of data that the ground surface is flat, such that by removing a linear trend along each line of elevation data an estimate of the ground clearance will be obtained. Figure 9 demonstrates the effect of removing a linear trend from the elevation data for one of the Camp Sibert data cells. Figure 9(d) compares the measured data with the background response modeled by using the estimated ground clearance. The difference between modeled and observed data is likely due to inaccuracies in the ground clearance estimation process. We note that difference in modeled and observed response could also be due to the median filtering applied to the data.

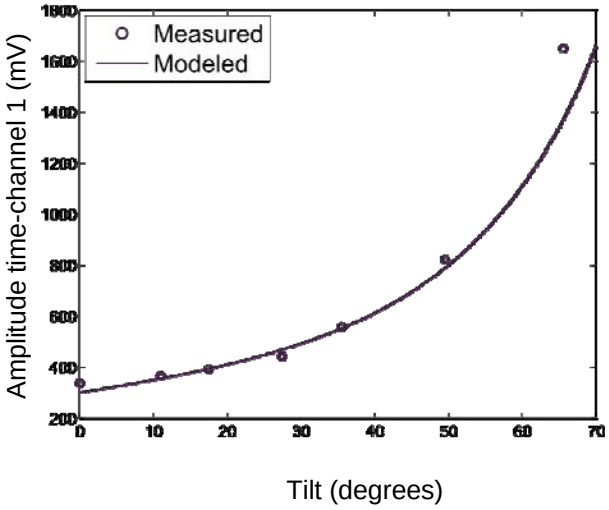
There are a number of factors that affect the accuracy of our processing:

- The MTADS data analyzed here were first leveled using a demedian filter. The demedian filter produced some artifacts in the data.
- The effect of topography was not included in the modeling. The same ruts and bumps on the ground surface that cause the MTADS trailer elevation to change can also produce EM anomalies. That is, a void beneath a sensor results in a decrease in amplitude while a bump produces an increase in the secondary field amplitude.
- In order to reduce computation time, the primary fields generated by each of the EM61 transmitters were generated with only 4 dipoles.
- We assume the magnetic susceptibility within a cell is constant. Data features produced by small spatial scale variations in magnetic susceptibility are not modeled.

Thus, there are limitations to the procedures used (the lack of detailed topographic model is the most limiting factor) but nevertheless the results are quite revealing.

(a) Setup for the Tilt test

(b) Tilt test modeling result



(c) Setup for the height test

(d) Height test modeling result

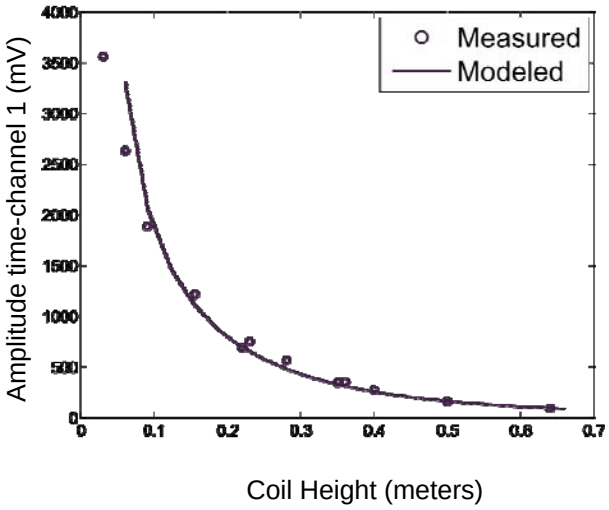


Figure 8: Tilt and height testing modeling results. A Geonics EM-61 Mark 2 sensor was used to collect data on Kaho'olawe, Island at a number of different heights and tilt angles. The signal was calculated by representing the transmitter loop with 8 dipoles.

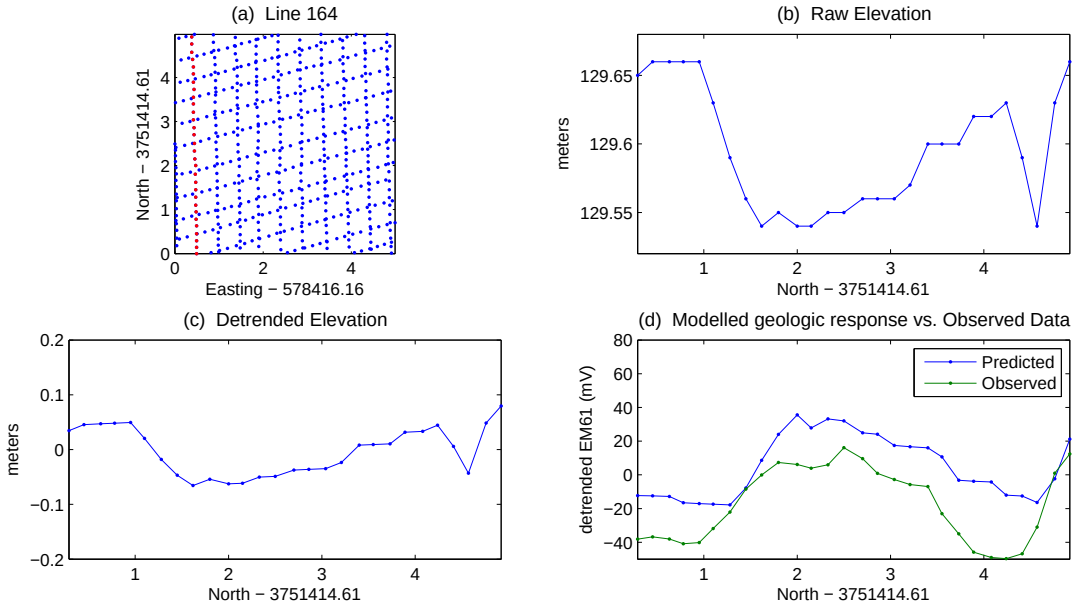


Figure 9: Example of how detrending the elevation can produce inaccurate results: (a) plots the line of interest from Cell 809; (b) the raw elevation; (c) The detrended elevation which was obtained by removing a linear trend. (d) the data and modeled background;

3.2 Results

The MTADS data were acquired along North-South and East-West lines. Many of the geology related anomalies on the site were due to ruts that ran along an East-West direction. When the MTADS trailer wheels would pass over these ruts, the distance between the magnetic soil and the sensors would decrease resulting in an increase in the signal measured by the sensors. These cross-track ruts were not as pervasive during E-W traverses of the site. For data collected along the East-West direction, the sensor motion anomalies were generally much smaller in amplitude. An example of this type of anomaly was found in Cell 644 (Figures 10-12). Figure 10(a) plots the first time channel of de-trended data. In the N-S lines there is a maximum anomaly magnitude of 40 mV, while the maximum amplitude in the E-W lines is 10 mV. Figure 10(b) plots the elevation determined from the three GPS sensors. Figure 10(c) plots the ground clearance estimated by removing linear trends from the elevation data. Removal of a linear trend predicts that the ground clearance varies by about 13 cm for data collected along N-S lines, and 5 cm for data collected along E-W lines. By comparing Figures 10(a) and (b), we see that the regions where the MTADS ground clearance is predicted to be closer to the ground are correlated with higher amplitudes in the recorded signal. We have found that there may be a slight lag issue with the data, as we consistently achieve better correlation between the ground clearance data and the secondary field when lagging the position by a couple of points. Figure 10(d) plots the data predicted by modeling the background with a constant magnetic susceptibility. We see that general shape of the predicted data is similar to the measured data. Figure 11 and 12 compare the data as profiles along N-S and E-W lines, respectively. In both figures four lines over the center anomaly are plotted. It is clear that anomaly 644 is caused by an increased response from magnetic soils due to the being closer to the ground during the NS traverse. There are smaller scale geology signals in the EW lines since there are smaller variations in the ground clearance of the sensors.

We have already seen that we can effectively eliminate many geological anomalies through the significant differences in energy or amplitude between the N-S and E-W lines. The geological anomalies with significant energy along both directions present a bigger problem. There were twenty anomalies classified as geology that had significant energy in the E-W data in addition to having a ground clearance variation of greater than 10 cm. These anomalies can be divided into four groups

1. Sensor motion anomalies. For these anomalies, estimated ground clearance and the sensor data appear to be correlated. Most of these types of anomalies can be eliminated by comparing E-W and N-S amplitudes.
2. Sensor motion anomaly and compact target. For these anomalies, there is appears to be a compact target present as well as a portion of signal correlated with the estimated ground clearance. For these anomalies there is a significant difference in amplitude between the E-W and N-S data.
3. Larger scale geology. Larger linear features in the data that were not removed using the filter.
4. Compact anomaly in both N-S and E-W data. Sensor data not well correlated with estimated ground clearance. The compact anomaly could be due to buried metallic target that was not found during excavation or possibly due to a concentration of magnetic soil. Additional time decay information could be used to help determine the likelihood the anomaly was due to soil.

We now provide examples from each of the four groups.

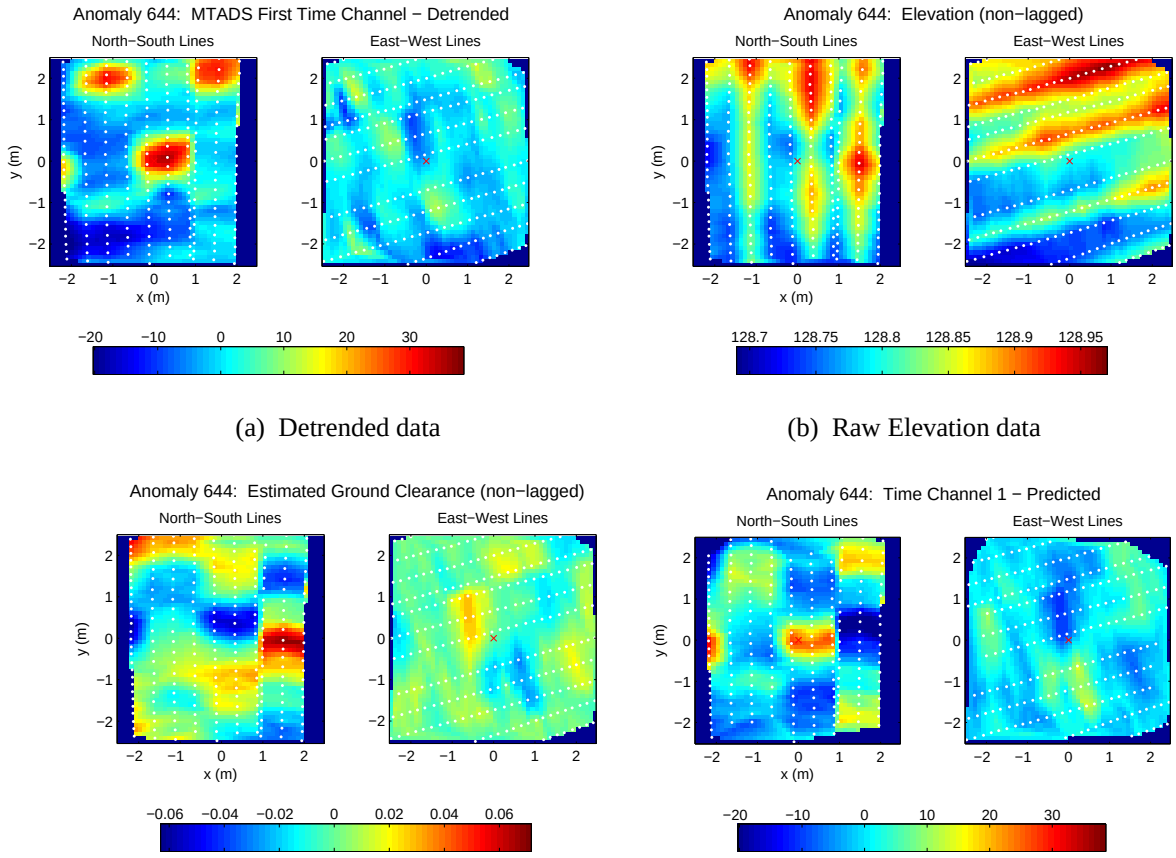
1. Sensor motion anomalies.

Figures 10-12 contained an example of a sensor motion anomaly (anomaly 644). In that case, the anomaly could be identified as originating from sensor motion by recognizing that the E-W lines have a low amplitude. If E-W data were not available, the predicted response based on a magnetic background would have indicated that the anomaly is likely due to sensor movement. Figures 13 and 14 contain results of modeling data from Cell 658. For this cell, there is a 60 mV anomaly in the E-W lines. There was poor data coverage in the N-S lines, such that comparison between N-S and E-W data magnitudes is not possible. The estimated ground clearance (Figure 13c) indicates that there may have been a depression that caused all of the sensors to move closer to the surface. The range of ground clearance is greater than 20 cm for this cell. Our modeling is able to predict the anomaly in the E-W data observed at the center of the cell. There are anomalies on the N-S line at an easting of ~ 1.5 m that are not modeled. However, these anomalies in the N-S are also seen in the E-W data suggesting that the signal is due to a compact target such as a piece of metal or a mound of magnetic soil.

2. Sensor motion and compact target anomaly

Several anomalies in the MTADS were classified as “can’t analyze” due to large differences in the anomaly magnitudes of the data in the N-S and E-W direction that could not be modeled using a dipole. In some cases, the difference in the magnitude could be partially explained by the EM response of the background geology.

Figure 15 contain data from anomaly 653. In the N-S data there is an anomaly of approximately 45 mV, while there is a 25 mV anomaly in the E-W data. Although the changes in ground clearance are relatively small (approximately 7 cm total variation in the ground clearance in the N-S data), modeling the background response predicts a geologic anomaly in the N-S data that has an amplitude of approximately 15 mV. This additional signal can partially explain discrepancy between the N-S and E-W data.



(a) Detrended data

(b) Raw Elevation data

(c) Ground Clearance estimated from elevation data

(d) Predicted data

Figure 10: Plan view of data from anomaly 644.

3. Larger scale geologic anomaly

Figure 16 contains an example of a larger scale geology feature that cannot be explained by ground clearance variations. The anomaly is present in both N-S and E-W data. The E-W data have a ground clearance variation of less than 5 cm. Due to the angle of the linear anomaly, it appears that the along line median filter was unable to eliminate this feature. Multi-time channel data would be useful for analyzing this type of anomaly.

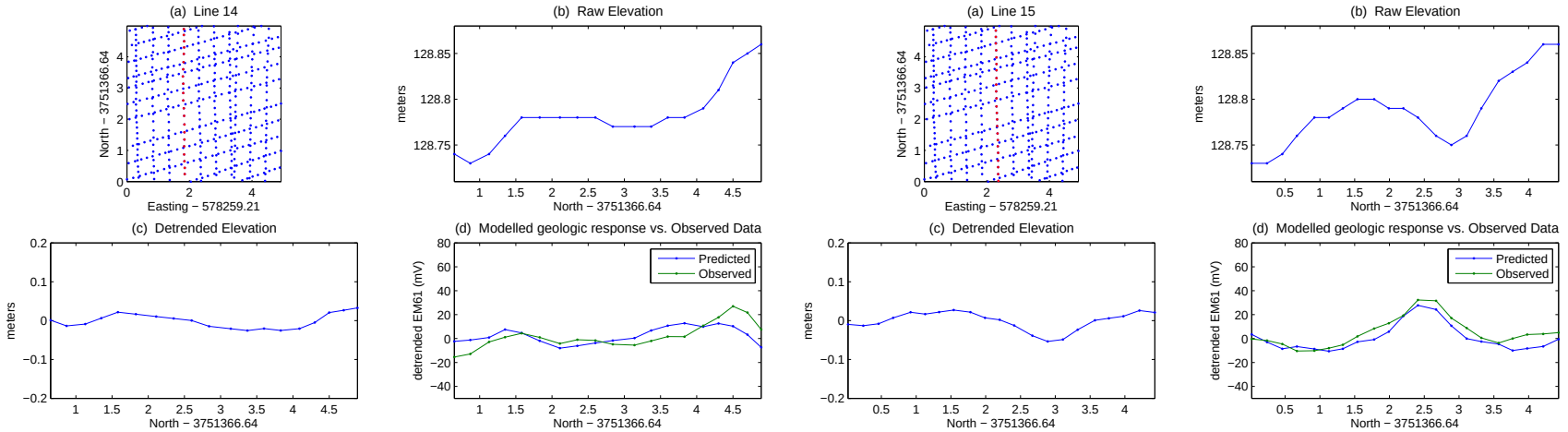
4. Compact anomaly in both NS and EW data

There are several targets in the data set that are classified as geology and have an anomaly in both the N-S and E-W data, where the magnitudes of the anomaly in both N-S and E-W are approximately equal. Two possible explanations for these types of anomaly are (1) there is a buried metallic target that was not found during excavation or (2) there is discrete concentration of magnetic soil. Figure 17 provides an example. Some of the differences in the shape of the anomaly in the NS and EW data could possible be due to sensor motion.

3.3 Conclusions

In this short study we examined if features in MTADS EMI anomalies could be due to sensor movement relative to the magnetic background at Camp Sibert. The background response was modeled by estimating a ground clearance from the elevation data and assuming that the background magnetic susceptibility was uniform in each cell. The accuracy of our modeling was limited due to filtering artifacts in the observed data, the accuracy of the ground clearance estimate, and small scale topography (i.e. depressions and bumps on the surface that would affect the measured data). The anomalies due to geology were generally greater in magnitude when data were acquired along N-S lines.

In many cases we found that small scale anomalies can be predicted using our modeling techniques. This modeling technique has the potential to reduce false alarms due to sensor movement. In addition this modeling will improve discrimination since a more complete forward model of the data could be incorporated into processing and inversion routines. In order for this modeling technique to be useful in a practical setting, better estimates of the ground clearance are required (e.g. by using altimeters mounted on the front of the array). In addition, sensors with minimal drift should be used, such that the need for applying a median filter to remove long spatial wavelength anomalies in the data is eliminated. A method that accounts for small scale topography should also be developed and tested.



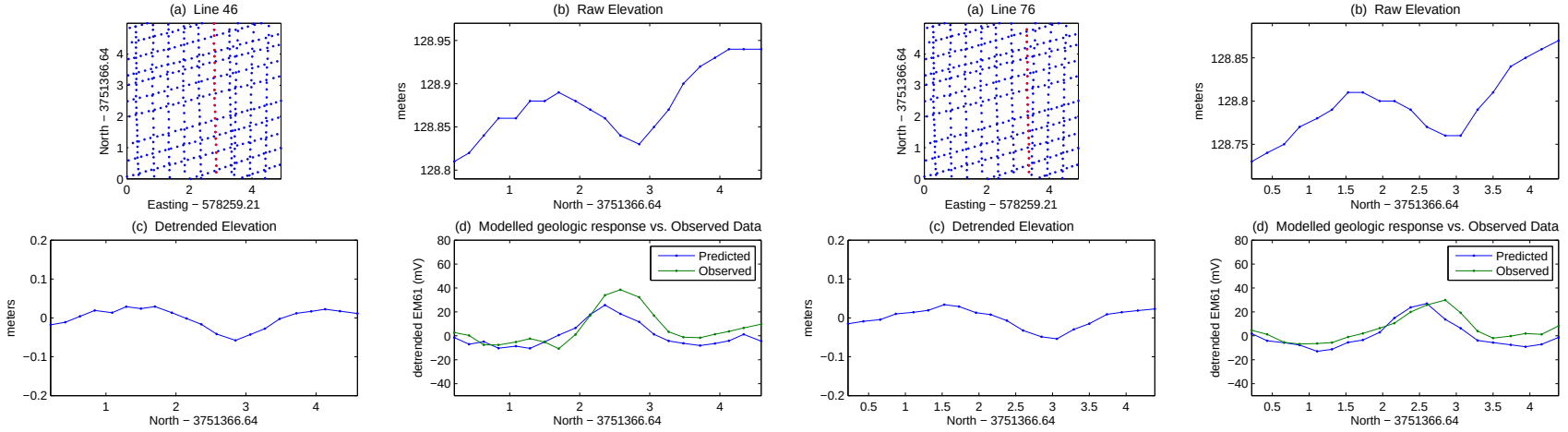


Figure 11: Anomaly 644. North-South Lines.

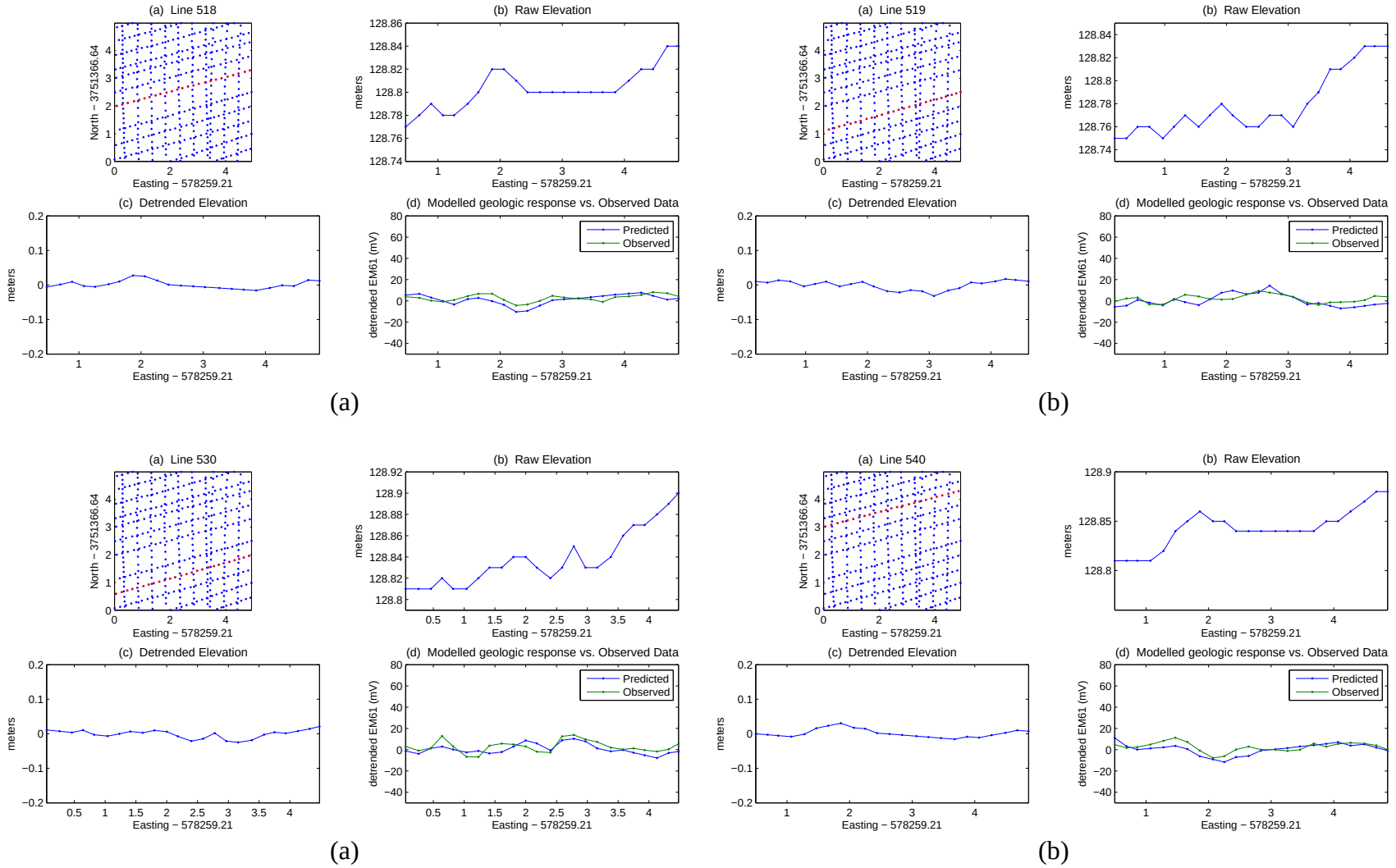
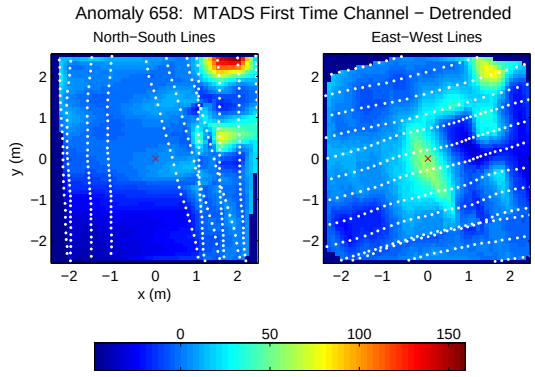
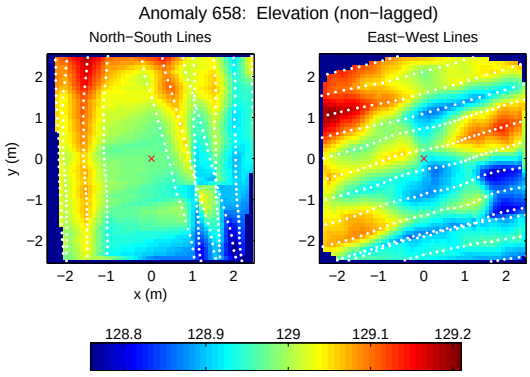


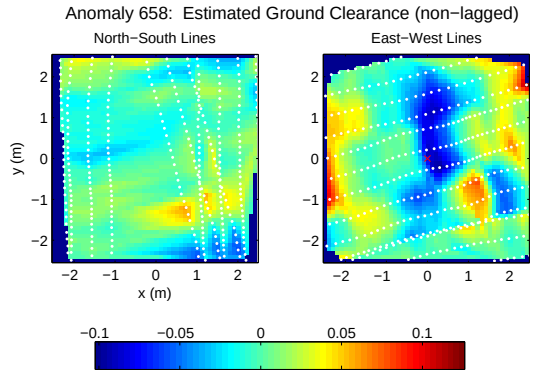
Figure 12: Anomaly 644. East-West Lines. Along EW lines variations in ground clearance are much smaller than along NS lines. In this case, even quite small variations in height (<5 cm) are correlated with sensor data.



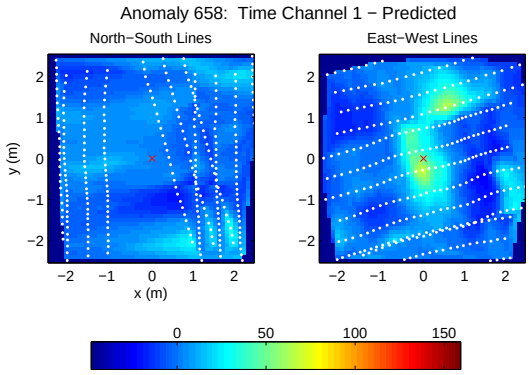
(a) Detrended data



(b) Elevation recorded by GPS

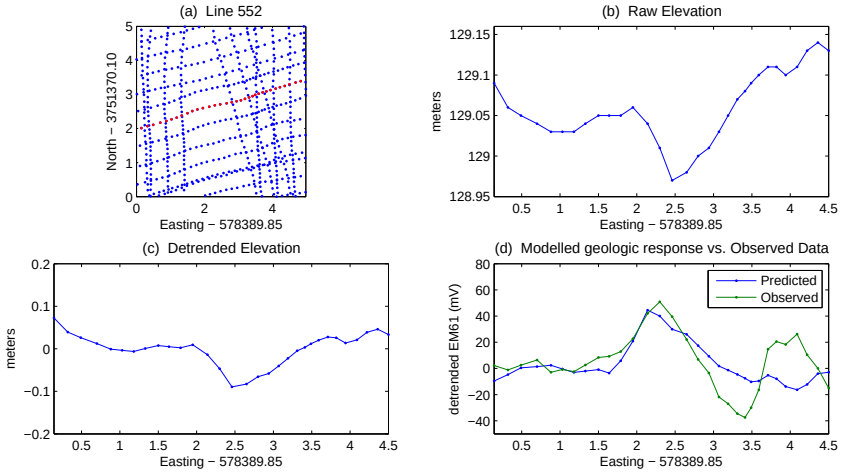


(a) Estimated clearance

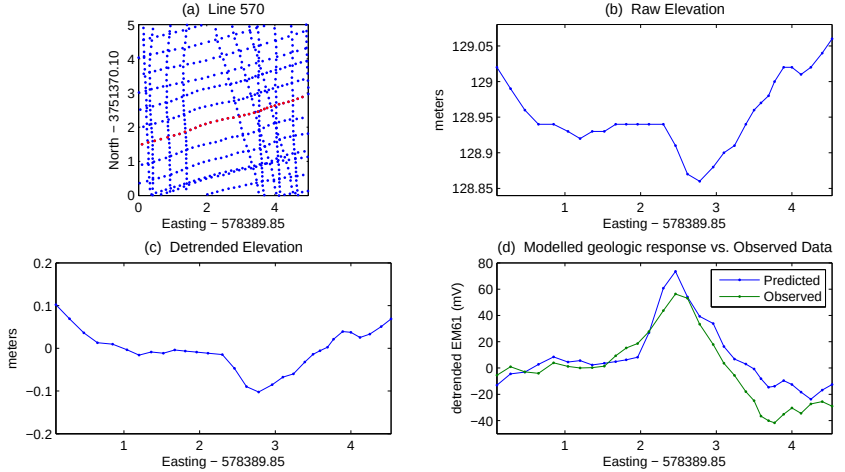


(d) Predicted data

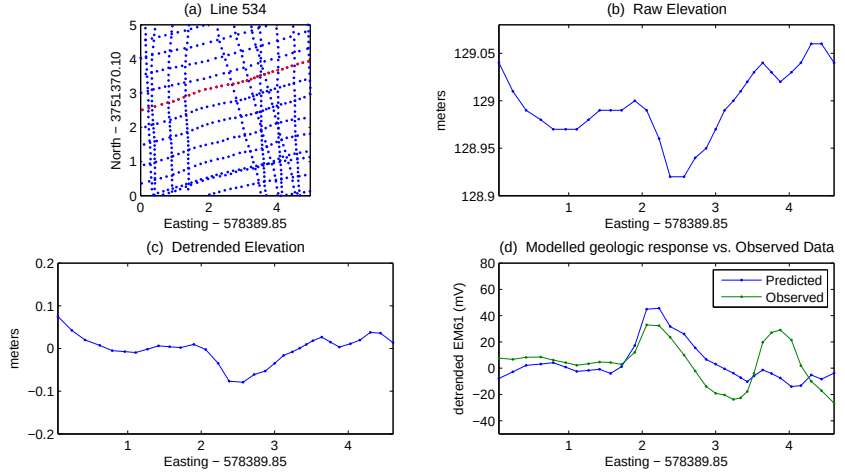
Figure 13: Plan view of anomaly 658.



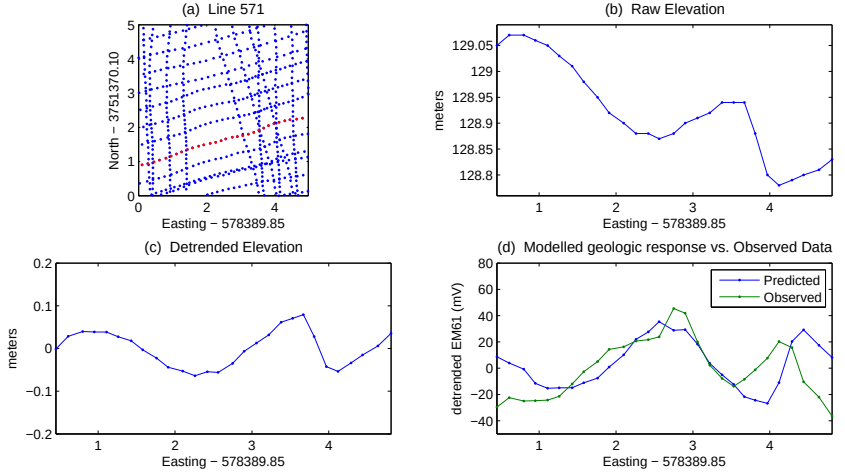
(a) Line 552



(b) Line 570



(c) Line 552



(d) Line 571

Figure 14: Profile view of anomaly 658.

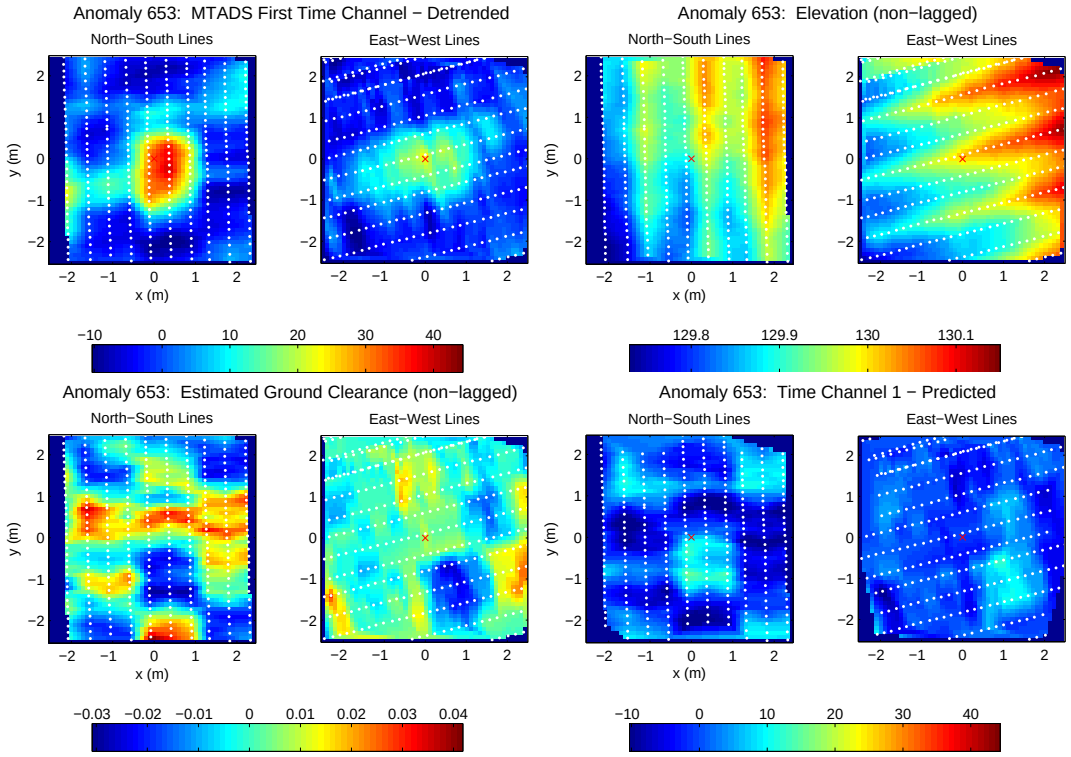


Figure 15: Anomaly 653, which has a compact anomaly in both N-S and E-W lines.

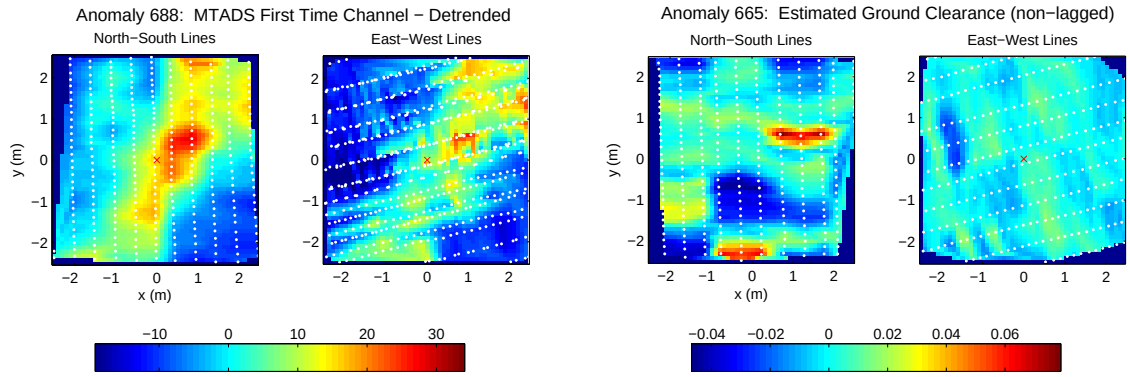


Figure 16: Plan view of anomaly 688.

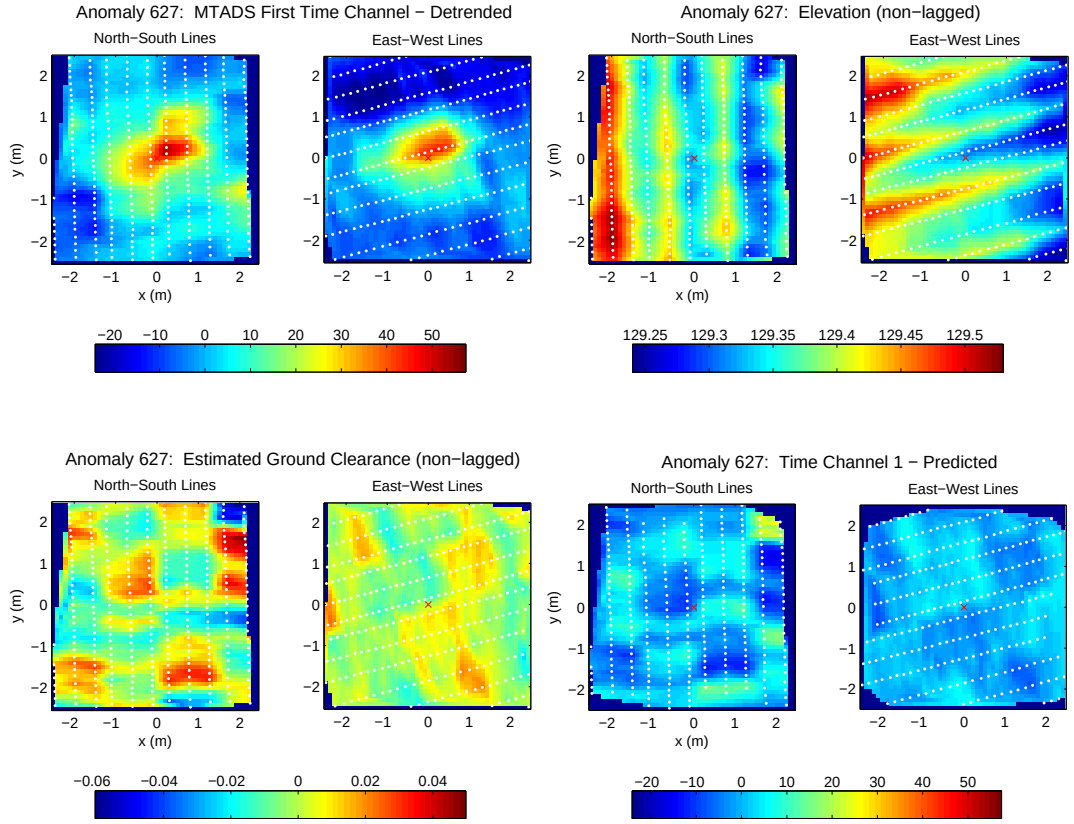


Figure 17: Anomaly 627 which has a compact target visible in both sets of profile data. Some of the difference in shape between the two anomalies can be attributed to movement of the sensor relative to the surface.

4 Determining the Receiver Operating Point in UXO discrimination

Most published UXO discrimination studies demonstrate the superior performance of a preferred algorithm with the receiver operating characteristic (ROC) or metrics derived from this curve. The ROC curve shows the proportion of true positives (UXOs in this context) as a function of the proportion of false positives (clutter). This curve is necessarily an uncertain estimate of the expected performance of a discrimination algorithm, since it is derived from a limited test set which is assumed to be representative of the underlying distributions of true and false positives. Accordingly, methods for computing ROC confidence intervals are increasingly applied in machine learning contexts [2], and to a limited extent in UXO work. For example, in [3] we used a bootstrapping technique to identify a discrimination algorithm which had the best expected performance given the observed training data. We measured algorithm performance using two metrics: the area under the ROC curve (AUC) and the false alarm rate (FAR). The latter quantity was defined by the "stop dig" criterion, that is, the number of false alarms required to find all true positives. The selection of this operating point is an important question in the application of statistical algorithms to UXO discrimination. In this work we address selection of the stop dig (aka operating) point, with particular focus on the electromagnetic datasets collected at Camp Sibert. A motivating example is shown in figure 18. This

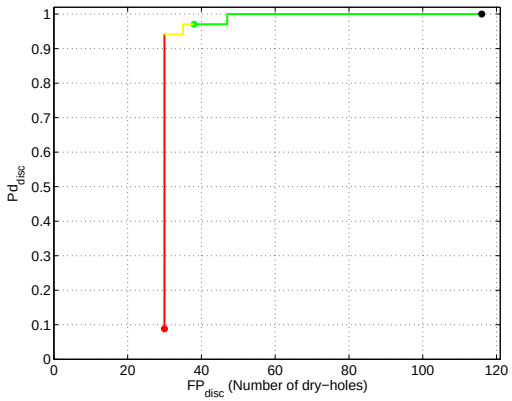


Figure 18: Receiver operating characteristic (ROC) generated by discrimination of 4.2" mortars at Camp Sibert, Alabama with EM63 data. The green circle represents the default operating point of the discrimination algorithm. Can't analyze anomalies are placed at the front of the dig-sheet and are not ordered.

curve was obtained by applying a probabilistic neural network (PNN) trained on features derived from EM63 data acquired at Camp Sibert. The objective was to distinguish the 4.2" mortars from metallic clutter, and the result in 18 was quite successful in this respect. However, one ordnance item remains in the ground at the chosen operating point. In the Camp Sibert demonstration, operating points were specified by expert judgement: i.e. we used visual inspection of the test and training feature vectors to decide when to stop digging. The failure of this approach in this example necessitates development of more objective methods for determining the operating point.

This remainder of this section of the report is organized as follows. In section 4.1 we consider the EM63 and EM61 features used for discrimination at Camp Sibert, Alabama. We show that the outlying UXO which are left in the ground in the ROC in figure 18 can be detected by incorporating feature vector uncertainty in the discrimination process. This obviates the need to select a more costly operating point later in the diglist, as in the original demonstration. In section 4.2 we discuss objective methods for selecting the operating point and demonstrate their application to the Camp

Sibert electromagnetic data. Finally in section 4.3 we discuss how to determine a domain in the feature space where classifier probabilities can be deemed meaningful.

4.1 Detecting outliers to the true positive distribution

Figure 19(a) shows training feature vectors extracted from geophysical prove-out data. The feature vectors were estimated from the observed data using a three-dipole model, with each dipole polarization decaying independently according to

$$L_i(t) = k_i t^{-\beta_i} \exp(-t/\gamma_i),\ i = 1, 2, 3. \tag{1}$$

The two parameters chosen for discrimination are the amplitude of the primary polarization, and the relative rate of decay of the primary polarization from the first to the 15^{th} time channel. There is near perfect separation between UXO and clutter classes in the training data. To generate classification predictions for these data we first applied a Probabilistic Neural Network (PNN), which represents each class as a superposition of Gaussian kernels and produces a nonlinear decision boundary. Applying this classifier to the test data shown in figure 19(b) with a default operating

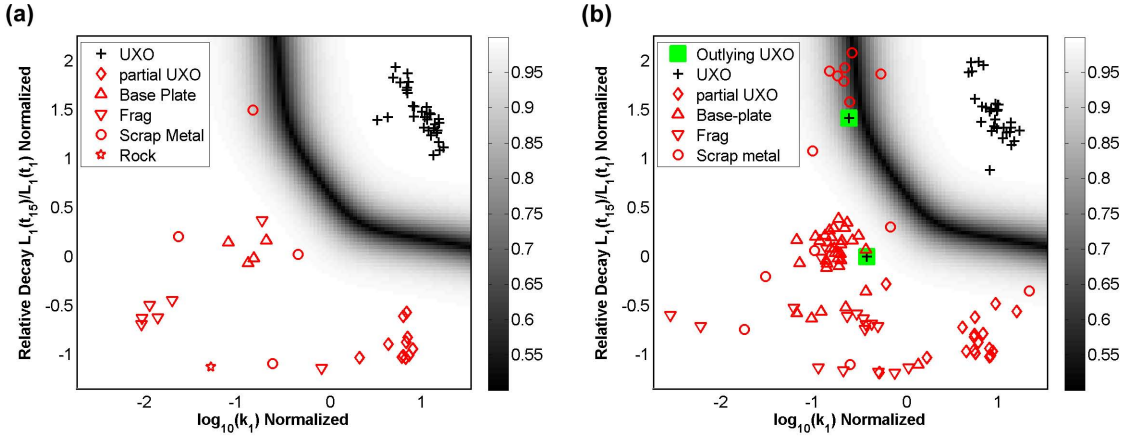


Figure 19: (a) Training feature vectors extracted from EM63 data. Grayscale image shows the maximum probability of membership computed for a probabilistic neural network. (b) Test feature vectors. Outlying ordnance feature vectors are highlighted.

point at the decision boundary defined by $Pr(TP|x) = 0.5$ misses two ordnance items in the test data [1] (highlighted vectors in figure 19(b)). Our failure to detect these outliers suggests that we have not properly characterized the distribution of UXOs in the feature space, and additional training vectors are required to augment our existing knowledge. However, we see in this case that, with the exception of the two outliers, all test UXO are readily found. It therefore seems reasonable to conclude instead that the feature vectors for these two items cannot be reliably used for discrimination.

Why has inversion failed for these particular targets? Figure 20(a) shows observed data corresponding to one of the outlying test feature vectors. We see that there is one line of data passing directly over the target which has a highly tortuous trajectory. This suggests poor positioning accuracy or real, but highly variable and rapid, motion of the sensor. Either possibility can result in a low SNR and compromise the fidelity of the estimated model. In particular, target depth becomes difficult to constrain when there is low signal to noise. The amplitude of the dipolar response is positively

[1] For the demonstration we set the boundary at 0.1 and only missed one of the UXO

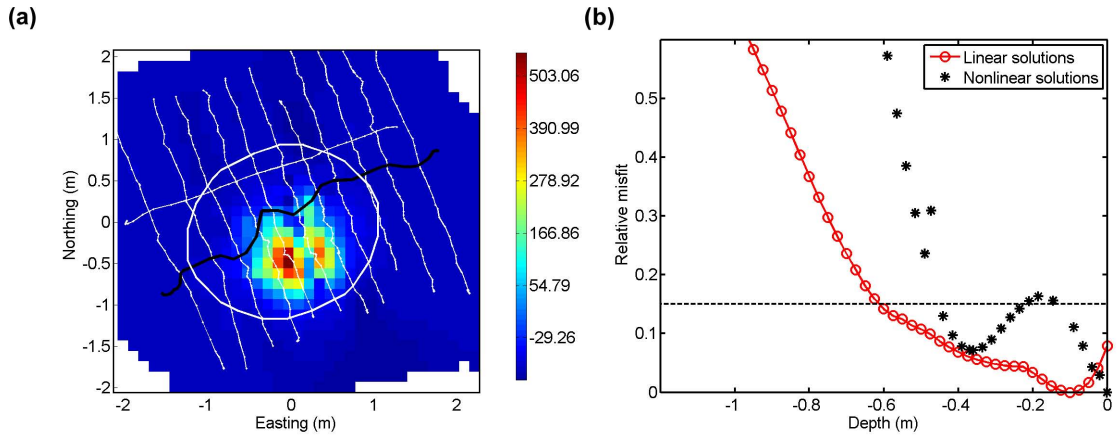


Figure 20: (a) Observed EM63 data at the first time channel over a 4.2" mortar. Highlighted line has anomalous tortuosity. Elliptical mask encloses data used for inversion. (b) Misfit vs. depth curve for inversion of data in (a).

correlated with depth, and so an incorrect depth translates to incorrect polarization parameters in equation 1. This problem has been successfully addressed by cooperative inversion of magnetic and electromagnetic data: constraints on target depth provided by inversion of magnetic data places the outliers in figure 19 squarely in the distribution of UXO [4]. If no magnetic data are available however, we must account for poor data quality in pre-processing or inversion. Data lines which have an anomalously tortuous trajectory can be easily detected with a pre-screener. For example, we might compare the length of the observed sensor trajectory with a smooth (e.g. straight line or spline) path. Lines with a large tortuosity can then be removed from the inversion. For the example shown in figure 20, this pre-screening technique easily identifies the problematic data and subsequent inversion produces a feature vector which is well within the distribution of ordnance.

However, positional errors are not always so obvious and we cannot rely solely on pre-screening to detect bad data. In [5] and [6], methods are developed to account for positional error in inversion using minimax and Bayesian formulations, respectively. Lhomme et al. developed metrics for measuring data quality in the figure of merit (FOM), which is an empirical measure of the expected reliability of the recovered model [7].

Here we exploit the misfit vs. depth curve also presented in [7] to characterize the uncertainty in the feature vector. This curve shows the dependence of the misfit function on target depth for multiple inversions. Two curves are generated: the first is the misfit vs. depth curve for solutions to a linear inverse problem assuming a fixed target location and solving for elements of the polarization tensor, the second shows solutions to the full nonlinear problem parameterized for target location and polarizations. The minimum misfit models from the linear problem are used to define starting models for the subsequent nonlinear inverse problem. Rather than displaying the absolute value of the misfits for converged models, we show misfits relative to the minimum misfit model.

Figure 20(b) shows the misfit versus depth curve obtained by inversion of the data in (a). We see that the converged models from nonlinear inversions trace out a misfit surface with multiple minima. The global minimum of this curve occurs at a depth of $z = 0$ and hence the estimated amplitude of the polarization (as represented by $\log_{10}(k_1)$ in 19(b)) is much smaller than expected for a 4.2" mortar. However, models residing in the local minimum at approximately 0.4 m depth lie well within the distribution of training UXO. This is shown in figure 21(a), which shows feature data for the same targets as in 19(b), except that all models with a relative misfit of 0.15 or less are displayed.

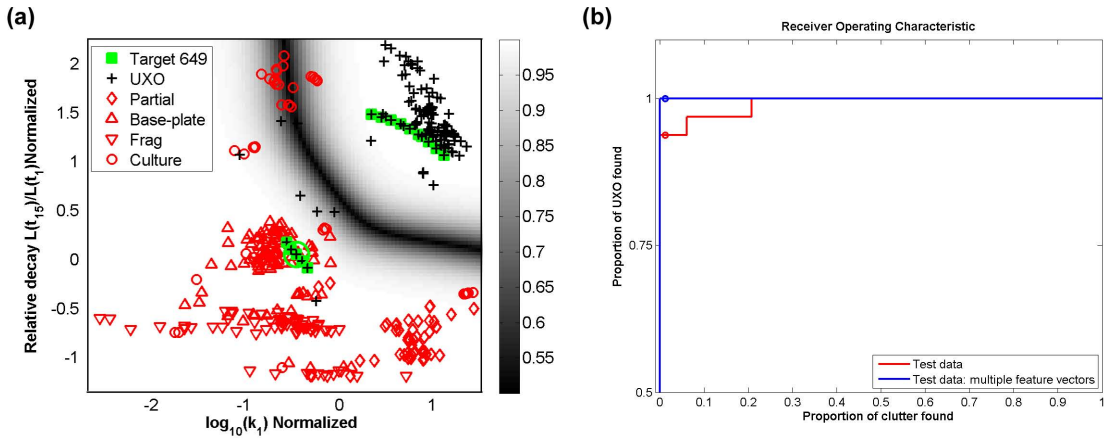


Figure 21: (a) Feature space including all feature vectors with 0.15 relative misfit or less. All feature vectors for the target in figure 20 are highlighted, and the minimum misfit feature vector is circled (b) Receiver operating curve generated by application of PNN classifier to data, using the minimum misfit feature vector or multiple feature vectors. The open circles on the ROC curves mark a classification boundary at $Pr(TP|x) = 0.5$.

For the example target considered here, the resulting set of feature vectors is drawn from the two local minima of observed misfit versus depth curve.

We can then use this ensemble of test feature vectors to generate an ordered list of targets. For targets represented by multiple vectors, we use the highest priority (most likely to be a UXO) vector to generate a prediction for that target. Applying this procedure to the test data produces a perfect ROC (figure 21): no clutter must be excavated in order to find all ordnance items. This obviates the need to set a more costly receiver operating point in order to detect outliers to the true positive distribution, as was done for these data in the initial demonstration [4]. While this method may help detect outliers to the true positive distribution, it may also increase the number of false positives excavated, especially when there is more overlap between true and false positives than in the EM63 data considered in this section. For example, if we apply the multiple feature vector idea to the EM61 cart data from Camp Sibert (figure 22), we are still able to find all buried ordnance, but more non-ordnance items must be excavated.

This last result raises an important question when applying this idea: how can we select the ensemble of models representing the range of possible solutions for a given target? If we set the misfit cutoff

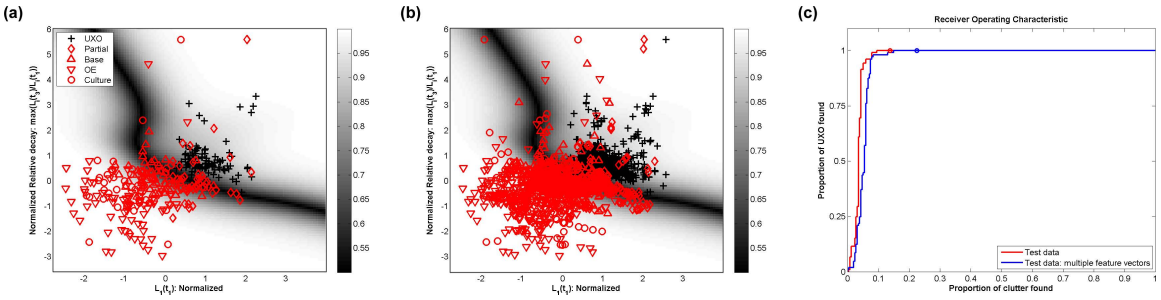


Figure 22: (a) Minimum misfit test feature vectors (b) Test feature vectors with 0.15 relative misfit or less. (c) Receiver operating curve generated by application of PNN classifier to test data, using the minimum misfit feature vectors or multiple feature vectors.

too low, then we may miss models which are closer to the "true" model (i.e. the model which best corresponds to the actual target depth). On the other hand, if we set the cutoff too high then we will increase the overlap between classes and degrade discrimination performance, as in figure 22. The selection of the relative misfit cutoff in these examples has been arbitrary: we use a default relative misfit threshold of 0.15 when selecting starting models from the linear misfit versus depth curve, and we have carried this parameter into the selection of models from the nonlinear misfit versus depth curve.

An alternative idea is to set the misfit cutoff based upon our knowledge of the probability distribution of the data misfit. Under the assumption of Gaussian noise, the data misfit function is a χ^2 distributed random variable with expected value $\mathrm{E}(\phi_\mathrm{d}) = N$, with N the number of data used in the inversion. This expected value is of particular use in underdetermined inversion: a recovered model with $\phi_d \approx N$ is deemed to adequately fit the data, providing an automatic method of selecting a regularization parameter (the discrepancy principle). In the case of parametric inversion the model has limited degrees of freedom to fit the data and selection of a regularization parameter by the discrepancy principle is usually not required. Indeed, when inverting real data it is not uncommon that the global minimum of the misfit does not satisfy $\phi_d \approx N$. The most likely explanation in this event is that we have incorrectly characterized the noise on the data, so that the normalized residual for a given datum may deviate from its expected value. To choose a relative misfit cutoff in the general case where the expected misfit is not obtained, we can try the following procedure:

1. At a chosen confidence level $1 - \alpha$, compute the confidence interval ([0,ϕ_α]) of the misfit. That is, integrate the χ^2 distribution with N degrees of freedom up to a critical value ϕ_α defining a probability $1 - \alpha$.

2. Compute the upper bound of the confidence interval relative to the expected value

$$\hat{\phi}_\alpha = \frac{\phi_\alpha - E(\phi_d)}{E(\phi_d)} \quad = \frac{\phi_\alpha - N}{N} \tag{2}$$

3. To account for errors in noise estimation, scale $\hat{\phi}_\alpha$ by a factor $\gamma = min(\phi_d/N, \gamma_{max})$, where ϕ_d is the minimum misfit of all converged models, and γ_{max} is chosen such that $\hat{\phi}_\alpha$ does not exceed a maximum value specified a priori.

Figure 23(a) shows the (unscaled) relative misfit as a function of N for $\alpha = 0.05$ and $\alpha = 0.01$ (0.95 and 0.99 confidence intervals, respectively). While the selection of the confidence level remains subjective, it is intuitively sensible to decrease the relative misfit cutoff as the number of data increases, as shown in figure 23. Large datasets are less influenced by outliers, so the potential for local minima is reduced and we can discard models with large relative misfit. Scaling the misfit cutoff (step 3 in the above procedure) is a necessary step. For example, if the minimum misfit significantly exceeds the expected value of $\gamma = 1$, then we have almost certainly underestimated the noise and consequently underestimated the uncertainty in the recovered model parameters. By increasing the misfit cutoff by γ we incorporate this uncertainty into the selection of an ensemble of feature vectors representing a single target. However, we find that this scaling can sometimes lead us to include high misfit models which increase the overlap between classes and greatly degrades the discrimination performance. Hence we restrict relative misfit computed in step 3 to a maximum value.

Applying this procedure to the test EM63 data with a 0.99 confidence level yields identical performance to the ROC shown in Figure 21 for multiple feature vectors. As shown in 23(b), a 0.95

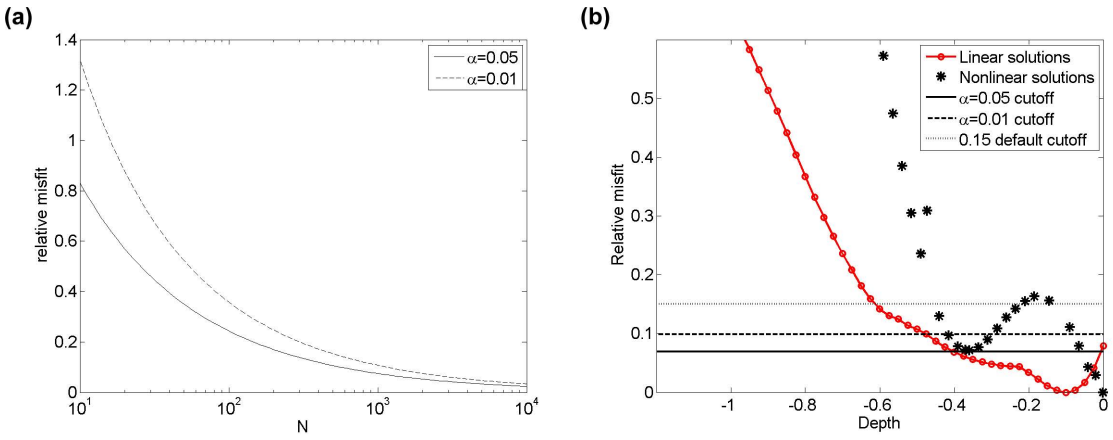


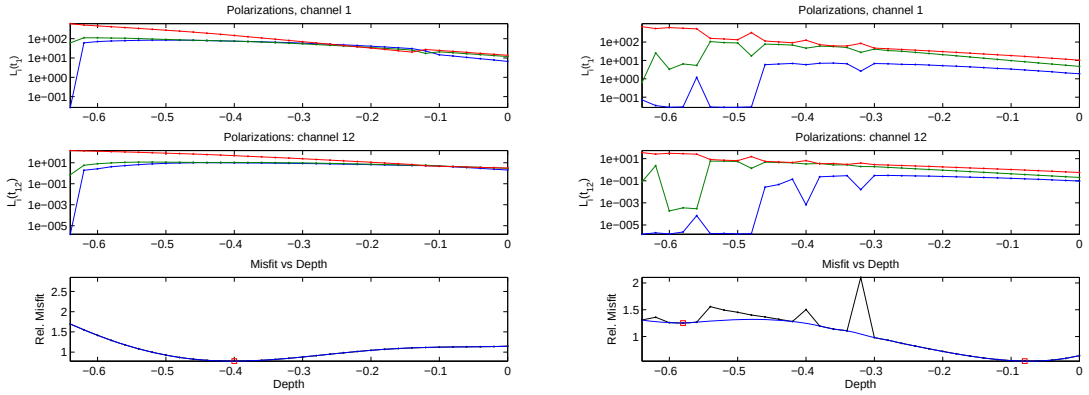
Figure 23: (a) Relative misfit cutoff as a function of N, the expected value of the misfit, for two confidence levels. (b) Misfit cutoffs computed for target in Figure 20.

confidence level results in a misfit cutoff which is too low for the target in Figure 20, and this target remains an outlier to the UXO distribution. For the EM61 data, a 99% confidence cutoff with γ_{max} specified to produce a 0.15 upper bound to the relative misfit produces similar performance to that shown in Figure 22. This is not surprising since 60% of the targets remain at the default 0.15 cutoff. The failure of this method to decrease the overlap between classes for these data suggests that while this procedure finds all ordnance at the 99% confidence level, it is not any better than using a fixed cutoff.

An alternative to using a fixed relative misfit cutoff to select models for discrimination is to identify local minima of the misfit versus depth curve. In the original demonstration we traced out the misfit versus depth curve for nonlinear models with only 10 models. To ensure that we have a sufficient number of points to pick off local minima, we have re-inverted the Camp Sibert EM63 and EM61 cart data with 41 starting models. Figure 24 shows example misfit versus depth curves for two fits to EM63 data, together with the dependence of the estimated polarizations on depth.

Ideally, our recovered misfit versus depth curve is a smooth function with one minimum. However, we find that some solutions do not converge, resulting in a curve with large outliers (e.g. Figure 24(a)). We have not yet been able to diagnose why the inverse code is unable to converge in some cases, and so we have used a despiking routine together with cubic spline interpolation to identify local minima of the misfit versus depth curve. Figure 25 shows the test data generated using this approach and the performance of the PNN classifier. There is a slight increase in the false alarm rate: a few clutter items now appear before all the ordnance have been excavated. However, this is a small price to pay to ensure that all ordnance are found. There is good potential to improve the method further through improvements to our procedures for estimation of the misfit versus depth curve and the algorithm for finding local minima.

The EM61 cart data has fewer time channels and less spatial coverage to constrain model fits. Consequently, a large number of targets in this data set have multiple local minima on the misfit versus depth curve. Many clutter items have a second minimum at depth corresponding to a large primary polarization. Hence discrimination with test data generated from local minima of the misfit versus depth curve (figure 26) is degraded significantly from the performance obtained when only the global minimum is retained.



(a) (b)

Figure 24: (a) Top: Estimated channel 1 polarizations as a function of depth. Primary, secondary and tertiary polarizations are shown as red, green and blue, respectively. Middle: Estimated channel 12 polarizations as a function of depth. Bottom: Misfit versus depth curve (black), cubic spline fit (blue) and local minima identified from the spline fit (red squares). Target is scrap metal at a depth of 15 cm. (b) As in (a), target is an ordnance item at 30 cm.

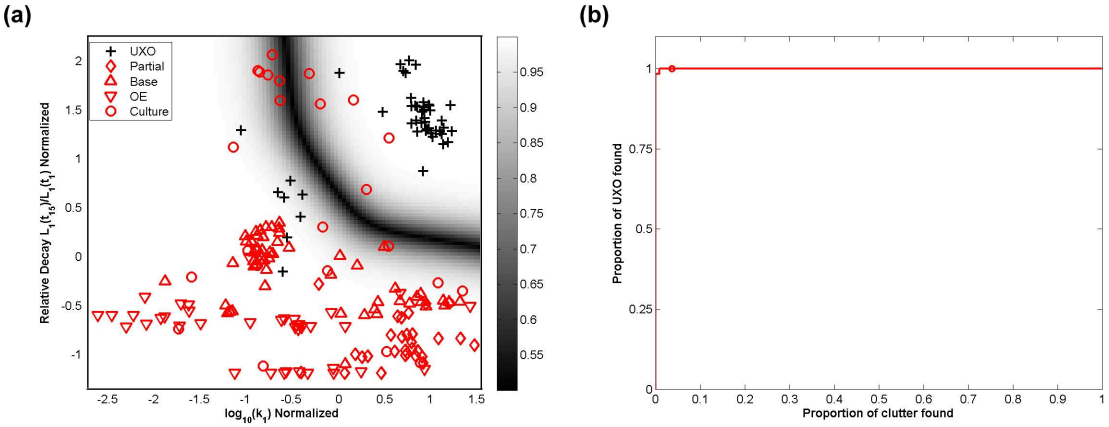


Figure 25: (a) EM63 test feature vectors selected from local minima of misfit versus depth curves. (b) ROC curve for EM63 test data in (a). Operating point is selected by bootstrapping technique presented in section 4.2.

4.2 Selecting the receiver operating point

In the previous section we investigated how multiple models can be used to characterize the variability of a feature vector for a target. In the case of the Camp Sibert EM63 data, this approach was able to detect all ordnance before digging a single clutter item. While eliminating the need for a more costly operating point, this result does not directly address the original task of selecting the receiver operating point. In this section we therefore review existing theory for choosing the operating point. We discuss why this theory is ill-suited for UXO discrimination and develop alternative approaches to selecting the operating point.

Following work in [8] and [9], let us denote the UXO distribution by T and the clutter distribution

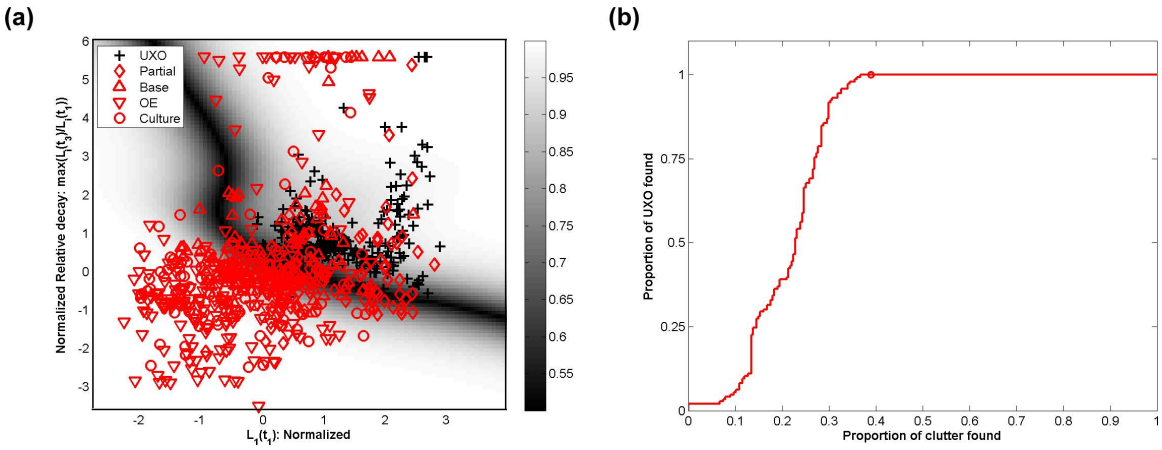


Figure 26: (a) EM61 test feature vectors selected from local minima of misfit versus depth curves. (b) ROC curve for EM61 test data in (a). Operating point is selected by bootstrapping technique presented in section 4.2.

as F. At a given threshold λ, we define the following probabilities

$$
\begin{aligned}
P(T|T,\lambda) &= \int_{-\infty}^{\lambda} p(x|T)dx \\
P(F|T,\lambda) &= \int_{\lambda}^{\infty} p(x|T)dx = 1 - P(T|T,\lambda) \\
P(T|F,\lambda) &= \int_{-\infty}^{\lambda} p(x|F)dx \\
P(F|F,\lambda) &= \int_{\lambda}^{\infty} p(x|F)dx = 1 - P(T|F,\lambda)
\end{aligned}
\tag{3}
$$

where $p(x|T)$ and $p(x|F)$ are the generating distributions of ordnance and clutter, as shown in figure 27(a). The receiver operating curve shows the variation of $P(T|T,\lambda)$ versus $P(F|T,\lambda)$ as λ is varied. The quantity $P(T|T,\lambda)$ is the probability that we correctly predict that an ordnance item is a member of the UXO class (a true positive), and $P(T|F,\lambda)$ is the probability that we incorrectly predict that a clutter item is ordnance (a false positive). These probabilities can be displayed as a "confusion matrix" (figure 27(b)). Confusion matrix probabilities alone are not sufficient to select an optimal λ, we must assign some cost C to each of the decisions we make. We denote the cost of a true positive as $C(T|T)$, and a false positive as $C(T|F)$. The costs of the other elements of the confusion matrix are similarly defined.

A notable feature of the ROC is that it is independent of a change in the relative frequencies of the two classes (the prior probabilities $P(T)$ and $P(F)$). However, once we assign a cost to each of the decisions in the confusion matrix, the prior probabilities become crucial because they determine how often we expect to incur the specified costs. Now given our confusion matrix probabilities,

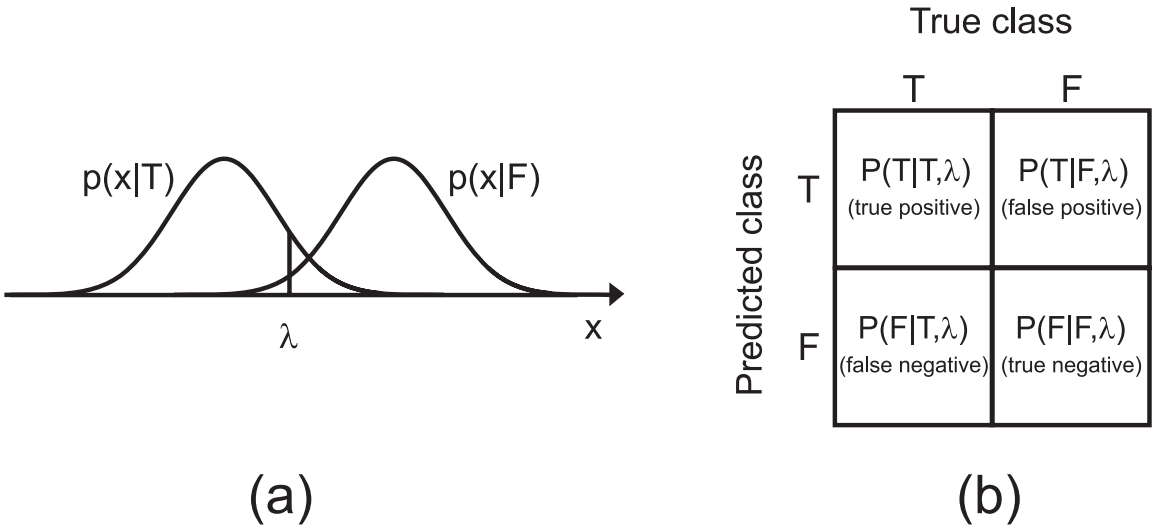


Figure 27: (a) The ROC is generated by integrating the distributions of true and false positives up to a threshold λ. This produces a confusion matrix of probabilities shown in (b), and the ROC is a plot of the first row of this matrix as a function of λ.

specified costs and prior probabilities, the expected cost (also called risk) at threshold λ will be

$$\begin{aligned} E(C(\lambda)) &= P(T|T,\lambda)C(T|T)P(T) + P(F|T,\lambda)C(F|T)P(T) \\ &+ P(F|F,\lambda)C(F|F)P(F) + P(T|F,\lambda)C(T|F)P(F) \\ &= P(T|T,\lambda)C(T|T)P(T) + (1 - P(T|T,\lambda))C(F|T)P(T) \\ &+ (1 - P(F|T,\lambda))C(F|F)P(F) + P(T|F,\lambda)C(T|F)P(F) \end{aligned} \tag{4}$$

Minimizing the expected cost with respect to λ gives

$$\frac{\partial P(T|T,\lambda)}{\partial P(T|F,\lambda)} = \frac{(C(F|F) - C(T|F))P(F)}{(C(T|T) - C(F|T))P(T)}. \tag{5}$$

Recalling that the ROC is a plot of $P(T|T,\lambda)$ versus $P(T|F,\lambda)$, the optimal operating point is defined by the point where the slope of the curve equals the right hand side of the above expression [8].

There are obvious difficulties with this approach to selecting the operating point. First, while in some applications (e.g. economics) costs are available, in UXO discrimination it is not readily apparent how to quantify the cost of leaving an ordnance item in the ground. We might encode expert, but subjective, judgements into costs. Intuitively, $C(F|T)$, the cost of mistakenly leaving ordnance in the ground, is the most costly mistake we can make, while correct decisions, $C(T|T)$ and $C(F|F)$, will incur less cost. Variable costs of excavating and disposing targets might also be considered: for example deep targets are more time-consuming (and hence more expensive) than shallow targets.

Even if costs can be chosen in a manner that will satisfy regulators, we also require the prior probability for each class to apply equation 5. In UXO applications, the frequency of clutter is almost always much greater than that of ordnance, and so we might think to set $p(F) \gg p(T)$. This will decrease the influence of ordnance items on the expected cost, and so an operating point which is optimal in the sense of equation 4 may not require us to dig all ordnance items. Hence we are faced with the problem of specifying two functions (costs and priors) from limited training data. A further problem is that the empirical ROC generated from observed data is a piecewise constant curve with

infinite or zero slope, and so finding an operating point with slope satisfying equation 5 requires that we fit a smooth function to the empirical ROC.

A final difficulty with the risk minimization approach is the requirement for accurate predictions of class probabilities from the distributions $p(x|T)$ and $p(x|F)$. Discriminative classifiers such as the support vector machine do not output probabilities and so the above framework cannot be applied unless we adopt some method to convert SVM outputs to probabilities. Even when probabilities are available, they may not be properly calibrated to the test data. For example, a classifier may provide an ordering of targets which identifies all UXO before clutter, but if, as mentioned above, we use prior probabilities to scale predictions, the predicted probabilities for UXO may be very small. In [10], the authors develop a binning scheme whereby classifier probabilities are reassigned as the proportion of training targets in that bin. This produces improved accuracy (i.e. lower false alarm rate) when applied to large test sets. However, the resulting classifier probabilities are discretized and the choice of discretization will likely have an important effect on performance.

Selecting the operating point: a bootstrapping approach

Given the problems with optimizing expected cost enumerated above, we choose instead to develop alternative methods for selecting the operating point. Our goal is a simple technique which requires minimal tuning on the part of the analyst.

An intuitive definition of an optimal operating point is one which allows us to find all ordnance items with a minimum number of false alarms. If the "true" ROC is defined by integration of the generating distributions in figure 27(a), then we must dig all clutter items in order to find all ordnance, since there is always a finite probability of observing a sample from the ordnance distribution if we have not integrated to ∞. Of course, in practice we can never observe the true ROC, our empirical ROC is generated from a finite set of feature vectors. In [3] we showed that for finite samples an operating point which finds all ordnance items can be found. Figure 28 shows the dependence of the FAR on sample size for Gaussian generating distributions. As the sample size N increases, we expect to observe more outliers to the UXO and clutter distributions, and so the expected false alarm rate increases to its limiting value of one. Also shown is the area under the ROC curve (AUC), which has no dependence upon sample size (although the variance of estimates from finite samples does decrease with increased sample size). In [3], we found that the FAR was a more useful statistic for comparing discrimination performance because it is sensitive to the effect of outliers on the ROC curve. In contrast, the AUC is an average over the entire curve and hence is a less useful metric when outlier detection is important.

The dependence of the observed ROC for a finite sample size is addressed to some extent in [11]. They model the distributions of true and false positives output by a classifier as normal distributions with equal variance. Unbiased estimates of this variance (which depend on the numbers of true and false positives in the training data) are then used to compute a threshold corresponding to an equal probability of membership in the true and false positive distributions. The resulting expression depends only upon the numbers of UXO and clutter in the training data. Operating points selected by this method typically identify at most 90% of the test ordnance items in the published results [11]. The distributions of true and false positives output by classifiers are often not of equal variance [2], and so the assumptions underlying this method may not be valid.

Following on the work in [3], we propose to estimate the optimal operating point for a finite test data set by bootstrapping the training data. Instead of estimating the AUC or FAR by bootstrapping, we estimate the probability $P(FAR)$ at which the all ordnance items are found. For a given set

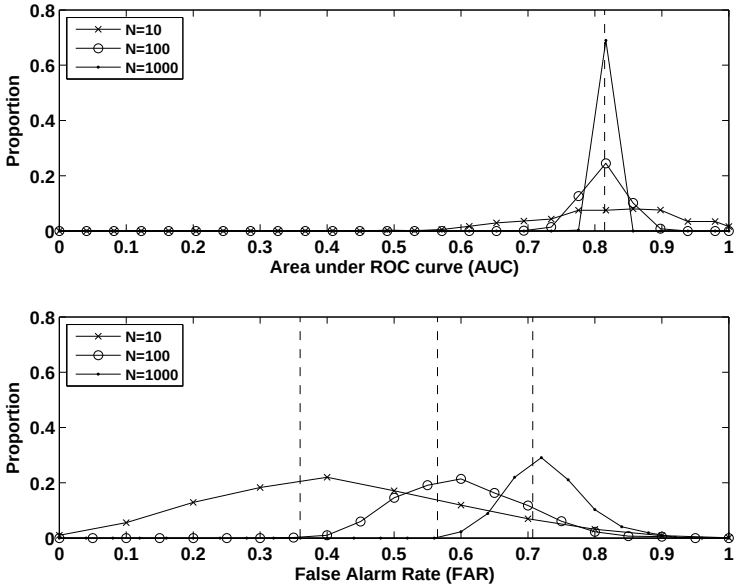


Figure 28: In each of 1000 trials, an equal number N of true and false positive samples were drawn from Gaussian generating distributions. The resulting distributions of estimated AUC and FAR are shown for varying sample size N. Vertical dashed lines show the expected AUC (upper plot, independent of N) and the expected minimum FAR for a sample of size N (lower plot).

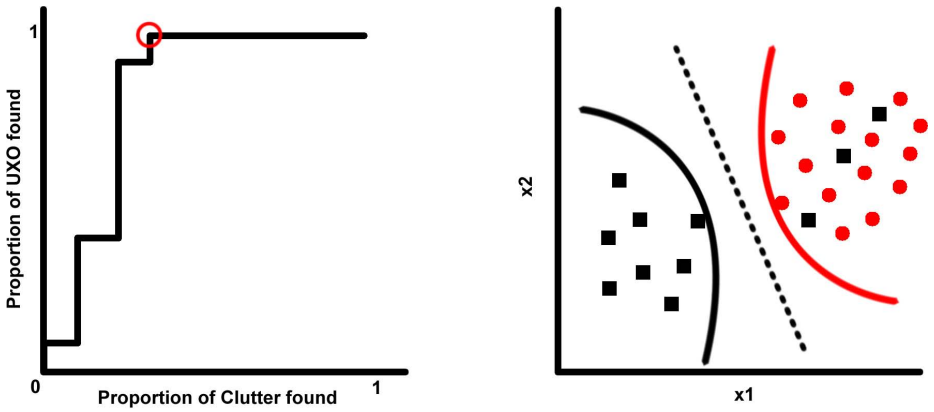


Figure 29: Computation of $P(FAR)$ from the ROC. Given an empirical ROC (left) generated from test data (right), we identify the point on the ROC where all ordnance are found (open circle). This corresponds to a probability contour (red line) in the feature space. The next target in the diglist (a clutter item) also defines a probability contour (solid black line). Then $P(FAR)$ is taken as the mean probability of the two contours (dashed black line).

of feature vectors, we compute $P(FAR)$ as the average of predicted probability for the last UXO item and the first subsequent clutter item in the ordered list of targets, as illustrated in figure 29. The concept is similar to that used with the support vector machine: we choose a decision boundary which maximizes the margin between support vectors. In this case the support vectors are the minimum probability UXO and the subsequent clutter item in the diglist. Using this method to compute $P(FAR)$, the bootstrap estimate is computed by

1. Generating a bootstrap realization of training and test sets by sampling with replacement from the full set of labelled data.
2. Training the discrimination algorithm on the bootstrap training set
3. Generating predictions for both the bootstrap training and test sets
4. Estimating the performance statistic ϕ (i.e. $P(FAR)$) of interest, again for both bootstrap training and test sets. For a given bootstrap realization B, this produces the estimates $\hat{\phi}^B_{test}$ and $\hat{\phi}^B_{train}$.
5. Averaging the bootstrap performance statistics according to

$$\hat{\phi}_{0.632} = 0.632\hat{\phi}^B_{test} + 0.368\hat{\phi}^B_{train}. \quad (6)$$

6. Repeating steps (1)-(5) to obtain a distribution for $\hat{\phi}_{0.632}$.

Intuitively, the weighting of training and test estimates in equation 6 corrects for the unequal sizes of bootstrap training and test sets and ensures that all labelled feature vectors are included in each estimate $\hat{\phi}_{0.632}$. We found in [3] that while bootstrap estimates can provide a reasonable ranking of classifiers, they tended to overestimate classification performance. When estimating an operating point by bootstrapping the training data, we prefer to err on the side of caution and replace the 0.632 estimator with one which takes the minimum of the bootstrapped test and training statistics

$$\hat{\phi}_{min} = min(\hat{\phi}^B_{test}, \hat{\phi}^B_{train}). \quad (7)$$

Figure 30 shows 500 bootstrapped test and training realizations of $P(FAR)$ for the EM63 training data, as well as the minimum for each realization. Because there is no overlap between UXO and clutter classes in the training data (see Figure 19(a)), the minimum $P(FAR)$ is consistently around 0.5. The mean of the minimum $P(FAR)$ over all realizations is 0.51 and yields the receiver operating point shown in (b). This yields identical performance to that obtained with a default 0.50 cutoff in Figure 21. Figure 31 shows the identical analysis for the EM61 cart data. In this case the bootstrapped choice of operating point (0.55) does decrease the FAR slightly relative to a default operating point (0.5). We also see more variability in the bootstrap estimates of the minimum $P(FAR)$ as a consequence of the increased overlap between classes.

The success of the bootstrap technique for these examples perhaps depends less on the method itself than on having reliable training and test feature vectors. If the training data is a representative sample from the distributions of ordnance and clutter, then we can expect the operating point selected by bootstrapping to work well for the training data. Furthermore, the concept of incorporating multiple feature vectors for each target in the test data ensures that the operating points chosen from the test data are able to find all ordnance. No technique for selecting the operating point (cost minimization, bootstrapping) will succeed if there are outliers to the ordnance distribution such as those encountered in the EM63 test data. However, the bootstrapping approach does address several shortcomings of risk minimization. In particular, we avoid the specification of somewhat nebulous costs and directly optimize the operating point for the task at hand, namely, to find all ordnance. The method also circumvents the need to compute accurate probabilities and can be easily applied to classifiers such as the SVM.

In figure 28, we showed that the theoretical FAR grows as sample size increases, and hence $P(FAR)$ decreases with increasing sample size. We have not accounted for this effect in selecting

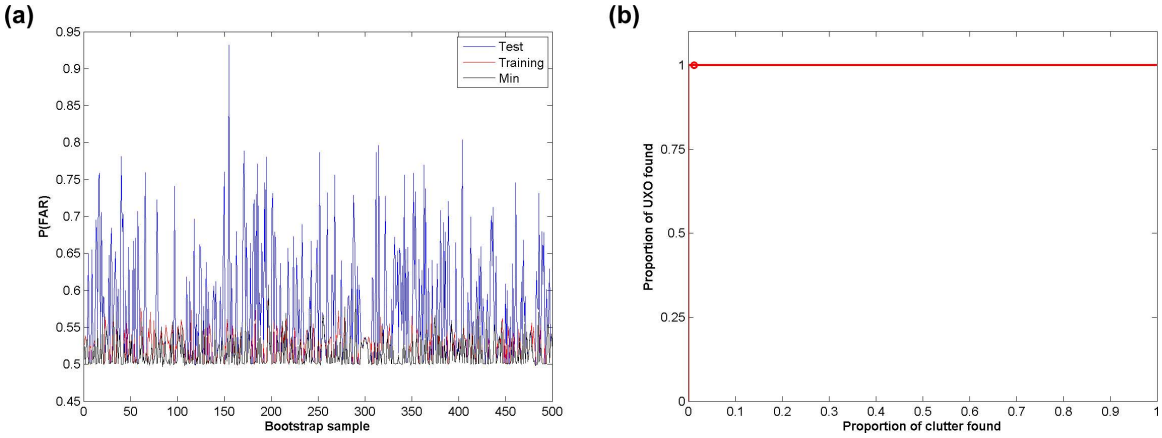


Figure 30: (a) Bootstrapped training and test realizations of $P(FAR)$ for EM63 training data. The minimum of the $P(FAR)$ for each realization is also shown. (b) ROC generated for EM63 test data with 0.15 relative misfit cutoff, and a $P(FAR)$ cutoff of 0.55, corresponding to the mean minimum $P(FAR)$ in (a).

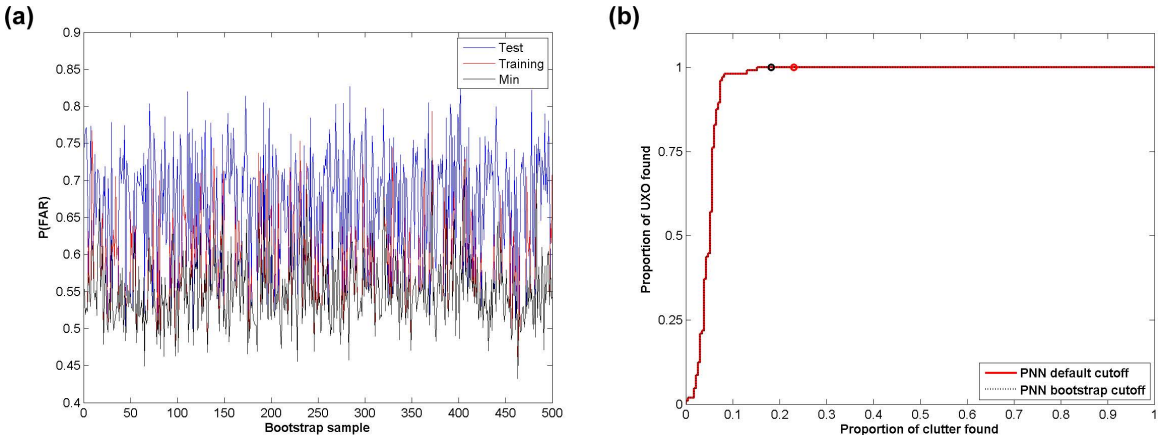


Figure 31: (a) Bootstrapped training and test realizations of $P(FAR)$ for EM61 training data. The minimum of the $P(FAR)$ for each realization is also shown. (b) ROC generated for EM61 test data with 0.15 relative misfit cutoff, and a $P(FAR)$ cutoff of 0.55, corresponding to the mean minimum $P(FAR)$ in (a). Also shown is a default operating point of 0.5 used in the original demonstration.

the operating point: the test data sets are substantially larger than the training data sets and so we expect more overlap between the two classes in the test data than is observed in the training data. To investigate this effect, we bootstrap increasing proportions of the EM63 and EM61 training sets. Figure 32 shows the dependence of the bootstrapped minimum $P(FAR)$ as a function of the proportion of training data used in the bootstrap. Interestingly, we see opposite trends in the EM63 and EM61 data. This is a consequence of the fact that there is no overlap between classes in the EM63 data. Regardless of the size of our bootstrap sample, we will not observe the expected decrease of $P(FAR)$ with increased sample size. Instead, $P(FAR)$ increases as probabilities are recalibrated with larger bootstrap training sets. For the EM61 data, however, there is overlap between the two classes and we do observe the expected dependence of $P(FAR)$ with increased sample size. This result indicates that we may wish to extrapolate $P(FAR)$ to a smaller value which is appropriate for a larger test set size. We might think to use the theoretical dependence of $P(FAR)$ on sample size for Gaussian distributions to make this extrapolation. However, one of the strengths of the bootstrap is that it makes no assumption about the distribution of the statistic of interest (i.e. it is a

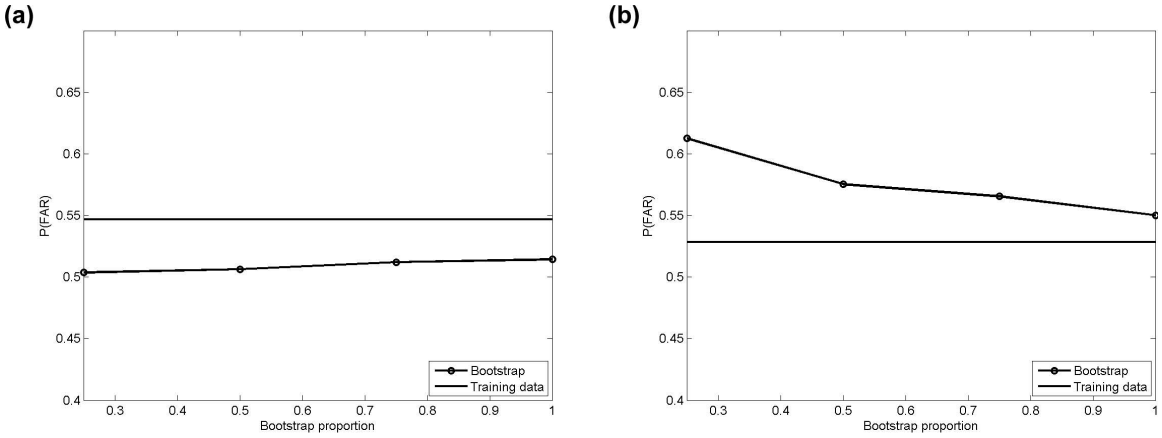


Figure 32: (a) Bootstrapped mean $P(FAR)$ as a function of increasing sample size for EM63 training data. (b) Bootstrapped mean $P(FAR)$ as a function of increasing sample size for EM61 training data. Solid lines indicate the $P(FAR)$ observed for the full training data set, without bootstrapping.

nonparametric estimate) and hence extrapolating with a Gaussian distributions would be inconsistent with the method developed here. This may be a moot point given the good performance of the bootstrapped operating points on the test data sets considered here, but an avenue for future study is to determine if and when the bootstrapped $P(FAR)$ breaks down for larger test sets.

Also notable in figure 32 is the difference between bootstrap estimates and the value obtained from the training data alone, without any bootstrapping. For the EM61 data we see that bootstrapping allows us to stop digging earlier than if we simply dig out to the last item in the training data. For the EM63 data bootstrapping forces us to dig slightly farther ($P(FAR) = 0.51$) than the training data requires ($P(FAR) = 0.55$), but there is no additional cost incurred when we apply this former threshold to the test data.

4.3 Ensuring a representative training set

As mentioned above, the bootstrapping technique requires that both test and training data sets are representative samples from the distributions of ordnance and clutter. The multiple feature vector concept ensures that our test vectors represent the range of possible solutions for a target. However, before applying our classifier and setting an operating point we must also be sure that our training data is reliable. If, for example, we include an outlier (corresponding to a poor fit) to the UXO class, then this outlier may dominate estimation of $P(FAR)$ and cause us to dig too far down in the test set diglist. This problem can likely be avoided in the majority of cases by careful quality control of fits. For example, when defining a training feature vector for a target, we should use the model which most closely corresponds to the actual depth of the target. For high SNR targets, this will usually correspond to the minimum misfit target. Even with careful QC of fits, we must still decide whether we have sufficient data from the GPO to train, bootstrap, and predict for the test data. In [11], a method is presented to build the training data set from scratch by iteratively identifying targets in the test data which are most informative, where information content is quantified with the Fisher information matrix. We hope to investigate the applicability of this technique to features derived from EM data in future work. In this report we investigate whether a representative training set can be built up using simpler methods.

Suppose we have a training set and a set of unlabelled test vectors. How can we ensure that a

classifier trained with our limited groundtruth will generalize to the test data and allow us to find all ordnance? If our model of the distributions of ordnance and clutter is correct, then the cumulative distributions of these classes in the test and training sets should be samples from the same model. We cannot generate the cumulative class distributions for the test data because we do not know the labels of the test vectors. However, we can generate the expected cumulative class distributions, as predicted by our classifier, as follows

1. Generate an ROC by thresholding on predictions (e.g. the predicted probability that a target is a UXO) for feature vectors. If multiple feature vectors are present for a target, use the maximum predicted probability for that target.

2. In a conventional ROC, we increment the count of items found by one as the probability threshold is decremented and we find a target belonging to the class of interest. Here we increment by the predicted probability, regardless of the class of the target (which will be unknown for test data). So, for example, a target which has a low probability of membership in the given class will contribute little to the total probability.

3. Normalize the total probability to one to obtain an expected cumulative distribution for a given class.

This procedure can be applied to both ordnance and clutter classes. For each class we generate an expected cumulative distribution from the test and training data. If both test and training sets are consistent with our model, then there should be no significant difference between the cumulative distributions. Figure 33 shows this procedure applied to EM63 training data and test data generated from local minima of the misfit versus depth curve. The ROCs in the top row of 33 illustrate our earlier remarks regarding calibration of generative classifier probabilities. Due to the normalization applied to probabilities, the predicted values are either nearly zero or nearly one. Consequently, the ROCs must be displayed logarithmically to see the differences between test and training cumulative distributions. Two-sample Kolmogorov-Smirnov tests at the 95% confidence level indicate that there is no significant difference between the test and training distributions of ordnance and clutter in the EM63 data. A similar result is obtained for the EM61 data. This may justify the use of bootstrapping to determine the receiver operating point. However, some investigation with synthetic data is still required to show that the cumulative distributions are significantly different when the test data is not a sample from the same generating distributions as the training data (i.e. our model does not apply to the test data).

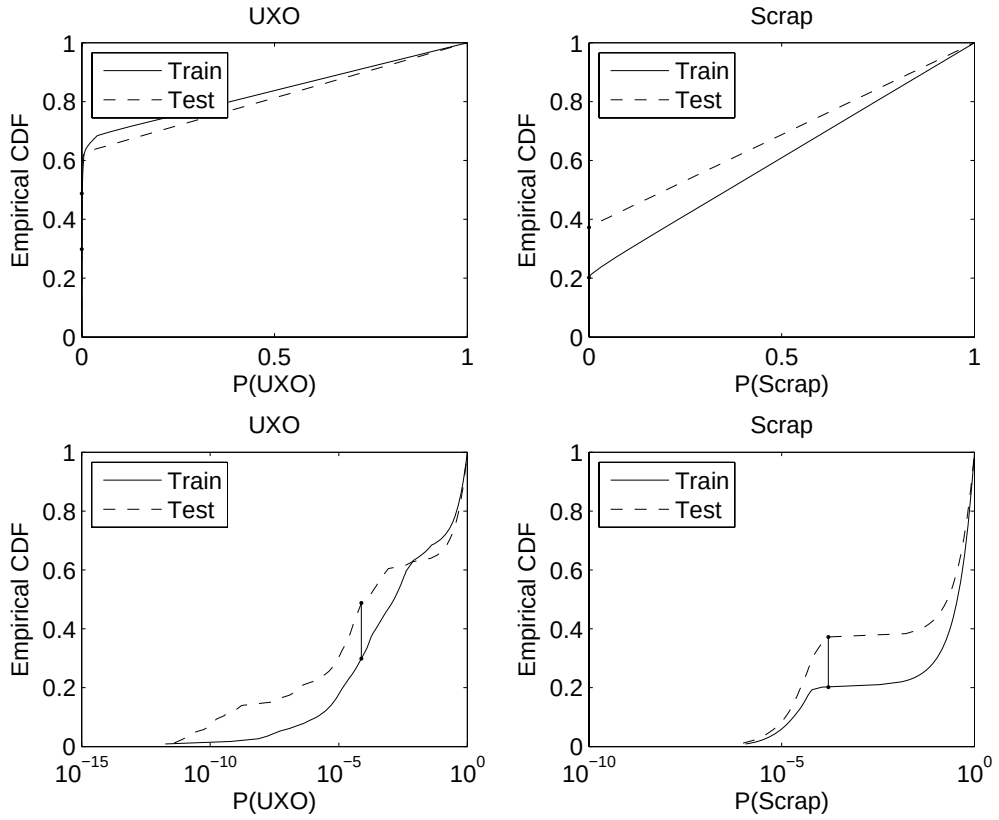


Figure 33: Top row: expected empirical cumulative distributions (CDF) of ordnance and scrap training and test vectors as a function of predicted probability output by PNN classifier. Bottom row: as in top, but with predicted probabilities (x axis) plotted logarithmically. The maximum difference between test and training distributions is plotted as a vertical line.

5 Re-inversion of EM63 data

During the process of developing the new classification strategies described in the last section, we noticed that there were some problems with the "zero" level of some of the EM-63 data (Figure 34). EM-63 "raw" data were obtained by removing a background level that was obtained through linear interpolation (as a function of time) between two static measurements conducted over the same location. Each static measurement provides an estimate of the EM-63 in-air response plus the response of the soil in the nearby vicinity. By repeating the measurement typically about a half-hour later, an estimate of any thermal drift in the EM-63 in-air response can be obtained, and removed (to within first order) through linear interpolation in time. The background-soil at Camp Sibert was not homogeneous, so this "zeroing" process often resulted in data with a residual soil signal: for instance, Figure 34b shows a profile of data over anomaly 647 (a 4.2" mortar that was placed in the "can't analyze" category) which has a residual soil response visible on the flanks of the main anomaly (the elevated response between fiducials 1 and 5 and 45 and 50). To remove this variable soil background, we detrend filtered the data using a 201 point moving median filter. This filter had the tendency to move the "zero" level to a negative value, so that the soil background became negative (Figure 34c). In most cases, the negative offsets were small and there was little influence on the recovered polarization tensor model. However, in some cases, the negative transients caused the polarization model to be placed closer to the surface, or caused the analyst to declare that the anomaly should be classed as "can't analyze" (this was the case for anomaly 647). We therefore applied a more sophisticated detrend filter that prevented negative transients (Figure 34d) and re-inverted all of the EM-63 data (without using depth constraints from the magnetometer).

Figure 35 compares the original ROC curve with a dig-list generated using the polarization tensor fits to the new detrend filtered data. The classifier and classification boundaries were identical to that used for the original dig-sheet. Two advantages of the re-inverted data are immediately apparent: firstly there are less can't analyze anomalies (10 compared to the original 30, although some of this could be attributed to a difference in analysts) and secondly, all 34 UXO are found before any clutter items are excavated. In contrast, 30 UXO were excavated before any clutter in the original dig-list, with two UXO listed as can't analyze, one recovered just before the operating point, and a forth that would have been a false-negative.

The performance of the EM-63 data when re-inverted is comparable to that of the cooperatively inverted EM-63 data. Using the magnetometer depth as a constraint, prevented the polarization tensor model from converging to a smaller, shallower depth solution on those anomalies were negative transients were a problem. By improving the detrend filter, negative transients were avoided, and the problem of convergence to smaller, shallower solutions was avoided.

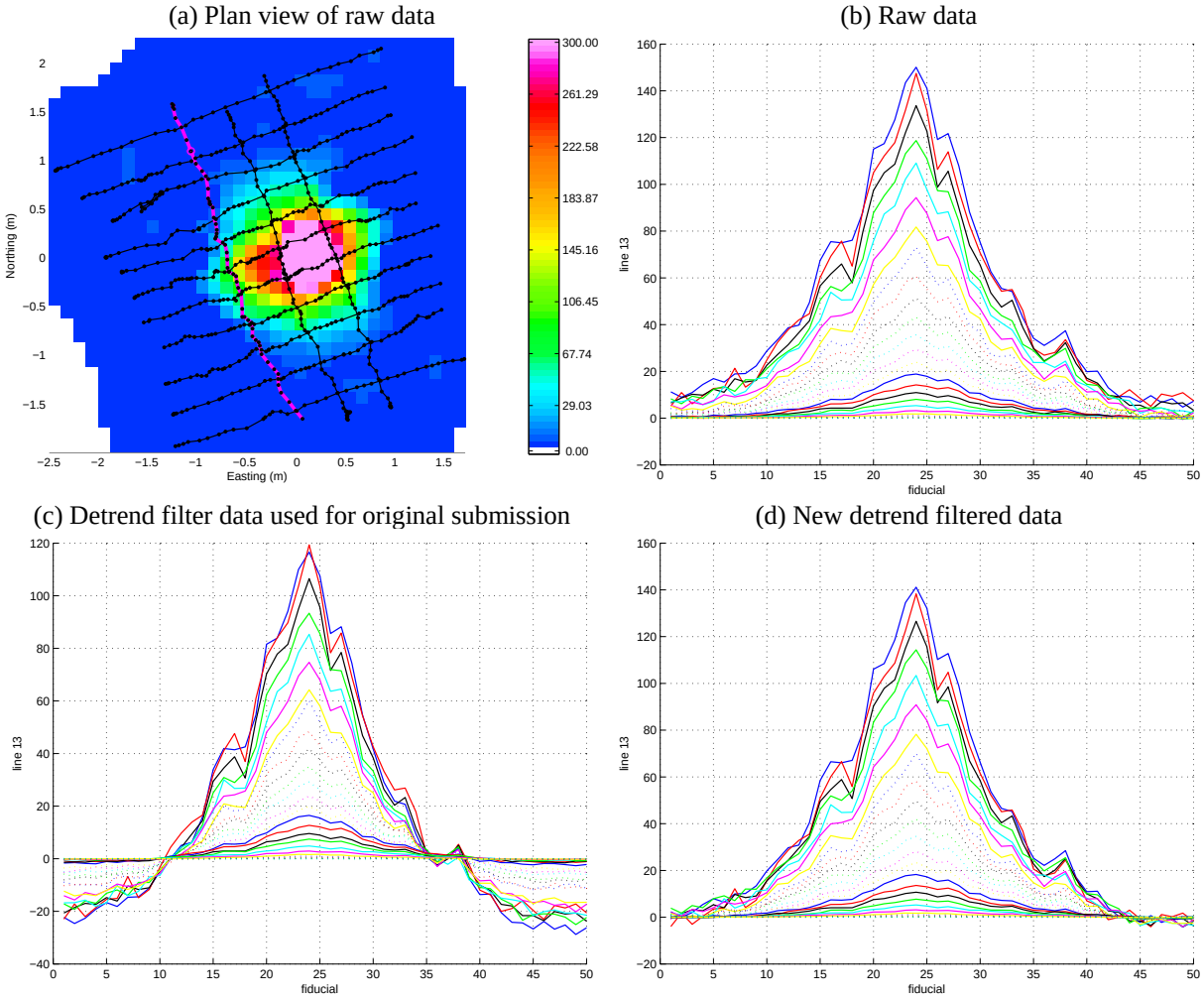


Figure 34: Plan view (a) and profiles (b to d) of anomaly 647, which was a UXO classified as "can't analyze" in the original submission: (a) Plan view with the location of the profiles in (b) to (d) marked as a pink line: (b) Profile of raw data (all 26 time channels); (c) Profile of the detrended data used in the original dig-sheet submission; and (d) Profile of new detrended data used to create a new dig-sheet.

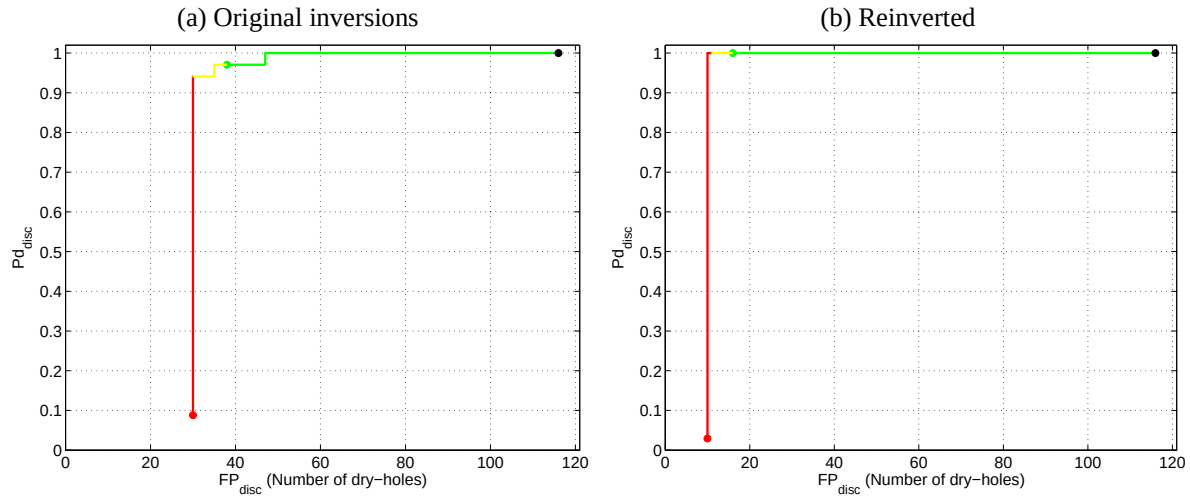


Figure 35: EM63 ROC curves of the original inversions (a) and after improving the drift correction (b).

6 Conclusions

In this report we further analyzed the EM-61 cart, MTADS EM-61 and EM-63 datasets collected at Camp Sibert. In particular we investigated methods to reduce the number of "can't analyze" anomalies in the EM-61 cart and MTADS EM-61 datasets, and methods for objectively setting the stop-digging threshold in all datasets.

The number of "can't analyze" anomalies in the MTADS EM-61 data could have been significantly reduced by monitoring the amplitudes or signal-to-noise ratio of data collected along North-South and/or East-West transects. Simply by rejecting anomalies that did not exceed the picking threshold of 25 mV on an E-W transect would have resulted in a reduction of 134 can't analyze anomalies in the MTADS EM-61 dataset and 88 can't analyze anomalies when the MTADS EM-61 data were cooperatively inverted. Similar reductions could be obtained by applying a metal-geology pre-screener which used the relative SNR in N-S and E-W to determine the likelihood of metal.

We investigated the cause of the can't analyze anomalies of geological origin whose amplitude exceeded 25 mV along the E-W transects. We hypothesized that many of these anomalies were due to sensor movement relative to ground. The background response was modeled by estimating a ground clearance from the elevation data and assuming that the background magnetic susceptibility was uniform in each cell. The accuracy of our modeling was limited due to filtering artifacts in the observed data, the accuracy of the ground clearance estimate, and small scale topography (i.e. depressions and bumps on the surface that would affect the measured data).

In many cases we found that small scale anomalies could be predicted using our modeling techniques and that sensor movement was indeed the likely cause. We also found that there were a number of metallic anomalies whose response could not be properly modeled due to variations in the signal from the background due to sensor movement. There were also several anomalies that were caused by longer wavelength spatial variations in magnetic soil properties that were not suppressed by the detrend filters that were used to pre-process the sensor data. We conclude that sensor movement relative to the ground is an important contributor to false-alarms and that we, in principal, have techniques to prevent such false declarations. However, the lack of good ground-clearance and micro-topographic information prevents the effective use of these techniques.

We also investigated discrimination performance when polarization tensor models were obtained from E-W transects only. When depths are constrained by cooperative inversion, there was very little difference in discrimination performance when using all lines or only East-West lines. With all data, only 2 false-positives were required before all 118 UXO were excavated compared to 4 false-positives for the East-West only data. When the cooperative constraints were not used there was one UXO that was recovered quite late in the dig-list, just 4 excavations before the operating point. Poor spatial coverage was deemed to be a possible cause for the poor fit obtained for that particular anomaly. We conclude, that, at least where data coverage was acceptable, there was very little benefit gained by collecting the MTADS data along perpendicular traverses.

At Camp Sibert, the stop-digging points were selected intuitively based on the characteristics of the training data. The thresholds were set very conservatively due to the potential for "outliers" (UXO with feature vectors that differ significantly from the training data). Here we attacked the outlier issue head-on by using multiple feature vectors for each anomaly. Performance was improved for the EM-63 (the false-negative was prevented) but not for the EM-61 cart data. The

performance of the EM-61 cart was degraded because many of the clutter items had relatively poor SNR and had larger, deeper, UXO-like models that fit the data relatively well. This caused the false-alarm rate to increase. The use of multiple feature vectors did prevent the occurrence of outliers and allowed us to objectively set the operating point based on a boot-strap analysis of the training data. For the EM-63 there was very little change in the operating point, while for the EM-61 cart, the multi-feature vector operating point could be set more aggressively. With this more aggressive cut-off the EM-61 cart performance was only slightly worse than we reported in the original demonstration report, with the added advantage that it was based on objective criteria. Any bootstrap analysis of the test-dataset will only be relevant to the training data if the test dataset is representative of the training dataset. Consequently, we developed a technique to determine the statistical similarity of the test and training datasets and used it show that both the EM61 and EM63 training datasets were representative of the test datasets.

References

[1] L.R. Pasion, N. Lhomme, K. Kingdon, S. Billings, D. W. Oldenburg, and J. Jacobson. Serdp 1573 annual report: Simultaneous inversion of unexploded ordnance and background geology parameters. Technical report, Strategic Environmetal Research and Development Program, 2007.

[2] Sofus A. Macskassy, Foster Provost, and Saharon Rosset. ROC confidence bands: An empirical evaluation. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[3] L. Beran and D. Oldenburg. Selecting a discrimination algorithm for unexploded ordnance remediation. *IEEE Trans. Geosci. Remote Sensing*, in press, 2008.

[4] S. Billings. Data Modeling, Feature Extraction, and Classification of Magnetic and EMI Data, ESTCP Discrimination Study, Camp Sibert, AL. Technical report, Environmental Security Technology Certification Program, 2008.

[5] Ashley B. Tarokh and Eric L. Miller. Subsurface sensing under sensor positional uncertainty. *IEEE Transactions on Geoscience and Remote Sensing*, 45:675–688, 2007.

[6] S. L. Tantum, Y. Li, and L. M. Collins. Bayesian mitigation of sensor position errors to improve unexploded ordnance detection. *IEEE Geoscience and Remote Sensing Letters*, 5:103– 107, 2008.

[7] Nicolas Lhomme, Doug Oldenburg, Leonard Pasion, David Sinex, and Stephen Billings. Assessing the quality of electromagnetic data for the discrimination of UXO using figures of merit. *Journal of Engineering and Environmental Geophysics*, 2007.

[8] Tapas Kanungo and Robert M. Haralack. Receiver operating characteristic curves and optimal Bayesian operating points. In *1995 International Conference on Image Processing (ICIP'95) - Volume 3*, 1995.

[9] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. Technical report, HP Labs, 2004.

[10] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[11] Y. Zhang, X. Liao, and L. Carin. Detection of buried targets via active selection of labeled data: Application to sensing subsurface UXO. *IEEE Trans. Geosci. Remote Sensing*, 42:2535–2543, 2004.